ETHICOMP 2021

Moving technology ethics at the forefront of society, organisations and governments

ETHICOMP BOOK SERIES

Edited by
Jorge Pelegrín-Borondo
Mario Arias-Oliva
Kiyoshi Murata
Ana María Lara Palma

UNIVERSIDAD DE LA RIOJA
Telefónica
Cátedra Telefónica. Aula Smartcities
UNIVERSIDAD COMPLUTENSE MADRID
UNIVERSITAT DE BARCELONA
UNIVERSITAT ROVIRA i VIRGILI

Edited by

Jorge Pelegrín-Borondo

Mario Arias-Oliva

Kiyoshi Murata

Ana María Lara Palma

ETHICOMP 2021

Moving technology ethics at the forefront of society, organisations and governments

ETHICOMP Book Series

UNIVERSIDAD DE LA RIOJA

Telefónica

Cátedra Telefónica. Aula Smartcities

UNIVERSIDAD COMPLUTENSE MADRID

UNIVERSITAT ROVIRA i VIRGILI

UNIVERSITAT DE BARCELONA

ETHICOMP BOOK SERIES

Title	Moving technology ethics at the forefront of society, organisations and governments
Edited by	Jorge Pelegrín-Borondo (University of La Rioja), Mario Arias-Oliva (Complutense University of Madrid), Kiyoshi Murata (Meiji University), Ana María Lara Palma (University of Burgos)
ISBN	978-84-09-28672-0
Local	Logroño, Spain
Date	2021
Publisher	Universidad de La Rioja

Publisher: Universidad de La Rioja, www.unirioja.es

Cover designed by Universidad de La Rioja, Servicio de Comunicación, and Antonio Pérez-Portabella.

ISBN 978-84-09-28672-0

To those who passed away due to the COVID-19 pandemic

The ETHICOMP Book series fosters an international community of scholars and technologists, including computer professionals and business professionals from industry who share their research, ideas and trends in the emerging technological society with regard to ethics. Information technologies are transforming our lives, becoming a key resource that makes our day to day activities inconceivable without their use. The degree of dependence on ICT is growing every day, making it necessary to reshape the ethical role of technology in order to balance society's 'techno-welfare' with the ethical use of technologies. Ethical paradigms should be adapted to societal needs, shifting from traditional non-technological ethical principles to ethical paradigms aligned with current challenges in the smart society.

Table of contents

1. Co-Creating Sustainable ICT Future Through Education

EDUCATIONAL SOFTWARE FOR SPEECH UNINTELLIGIBLE CHILDREN WITH DOWN SYNDROME

Katerina Zdravkova, Boban Joksimoski

University Ss. Cyril and Methodius, Faculty of Computer Science and Engineering (N. Macedonia)

katerina.zdarvkova@finki.ukim.mk; boban.joksimoski@finki.ukim.mk

INTRODUCTION

Down syndrome (DS) is a genetic disorder, which is associated with mild to severe intellectual disability and speech difficulties (Rice, 2005). Children with DS have atypical phenotypic features, including tongue anomalies, low oral-facial muscle tone, and difficulties in motor planning, resulting in severe acoustic alterations and disability of articulating some sounds (Carl, 2020). Heavy struggle with saying words and sounds was noticed among 85% of 1620 surveyed children with DS (Kumin, 2006). Being aware of their speech production deficit, they become frustrated and start using multimodal communication, combining manual signs, voice, facial expressions, and gestures (Toth, 2009). Unlike the problems with speech articulation, children with DS show excellent skills in gesture acquisition and production, compared to their peers with typical development (Deckers, 2017). Niki, the boy who successfully played the memory game (Zdravkova, 2020) and his mother managed to become proficient in Macedonian sign language (SL), establishing an easier mutual communication for the first time. This accomplishment raises the following questions: will the imposing of SL be beneficial for speech unintelligible children with DS; can educational games support understanding and acquisition of SL; how to design and assess children with DS who frequently do not master reading (Martin, 2009).

Enabling human rights, including the right to access information, to enjoy a decent life and express freely are vital to human welfare, dignity and active participation in community (UNICEF, 1989). Everyone is entitled to all these rights, embracing mentally or physically disabled, who are usually not self-reliable. Any focused, well-planned, and carefully implemented activity that improves the interaction with and among the speech unintelligible, without forcing them to leave their comfort zone and feel anxious due to the incompetence to adapt to new obligations can positively contribute to the improvement of the quality of life of these people. Since 1950s, many methods and assistive technologies intended to supplement or replace a wide spectrum of speech and language disabilities, have been developed as part of the augmentative and alternative communication (AAC) (Elsahar, 2019). Implementation of MAKATON, a sign language system that enables augmentative and alternative communication proved that their communication and socialization significantly improved (de Almeida Barbosa, 2018). All these arguments confirm that SL can be a valuable interaction alternative for speech unintelligible, including children with DS.

A plethora of applications are dedicated to SL interpreting. Popular Hand Talk Translator (www.handtalk.me), a 3D interpreter, which translates text and audio into American and Brazilian SL has already been used by half billion deaf and hard of hearing. Microsoft Translator (translator.microsoft.com) enables text-to-speech translation, and it has already been experienced by US students. It successfully converts raw spoken language and stutters into fluent American English. Microsoft also made similar feats, presenting a Kinect based system for SL translation (Chai, 2013). The Kinect sensor was later abandoned in favour of Intel Real Sense devices that provide similar features. Applications that focus on providing educational framework for SL are less popular than SL translators / interpreters. This is especially problematic when enabling learning of SL for children with intellectual

and speech disabilities. To assess the gamified education approach for children with DS, five popular applications in terms of downloads that focus on learning a SL have been assessed on several clear and distinguishable features. One aspect that is of interest for children with DS, is to create engaging environment and track the progress, usually by encouraging gamification features. Another important feature is the primary goal of the application, to enable an alternative way of communication or to establish it. The applications that rely on iconographic navigation (like ASL Fingerspelling game) are suitable for those children, who have no communication skills. The results are presented in Table 1 on the following page. They prove that educational games can significantly or partially improve interaction of both speech unintelligible and children with DS.

Table 1. SL acquisition applications and implemented approaches for visualization and learning.

Application	Availability	General features	Navigation and UI	Suitability for children with DS
SL ASL - Pocket Sign	Android, iOS	Video and image based, quizzes, progress tracking	Combination of text and iconography	Completely
ASL American Fingerspelling game	Android	Card and image-based content, quizzes, progress gamification	Combination of text and iconography	Completely
Sign Language: ASL Kids	iOS, Android	Video and image based, quizzes, progress tracking, gamification	Mainly iconographic	Completely
Hands On ASL Fingerspell with SL	Android, iOS	Learning fingerspelling, 3D hand models, quizzes and progress tracking	Text based	Partially
Mimix3D SL	Android	3D avatar, progress tracking, text to SL translation	Text based	Partially

DESIGNING EDUCATIONAL SOFTWARE FOR SIGN LANGUAGE ACQUISITION

Educational software for children with DS enabling acquisition of Macedonian SL is designed as a sequel of three games that complement each other. The goal of the first game is to demonstrate the sign language by enabling recognition and presentation of alphabets of Macedonian standard and sign language (Figure 1); the second aims to empower the acquisition of the most frequent words; the most advanced will support the creation of simple sentences.

Demonstration segment of the first game has already been created, following the nine recommendations of educational games for children with DS (Zdravkova, 2019). The application was created in HTML, powered by CSS framework for adding styles and colors, and React JavaScript library for user interface. It is currently available as a desktop application with optimized version for mobile phones and tablets. It presents SL alphabet with one hand (Figure 2), separate presentation of each letter (tab “Изучи ја азбуката”, Figure 3, left screen), matching of letters presented with SL and their written equivalent (tab “Погоди”, Figure 3, central screen) and a memory game with several levels (tab “Меморија”, simple level, Figure 3, right screen). All the written content on the screen is associated with a pre-recorded message with natural voice, bypassing illiteracy of many children with DS.

Figure 1. Home page of the Web application.

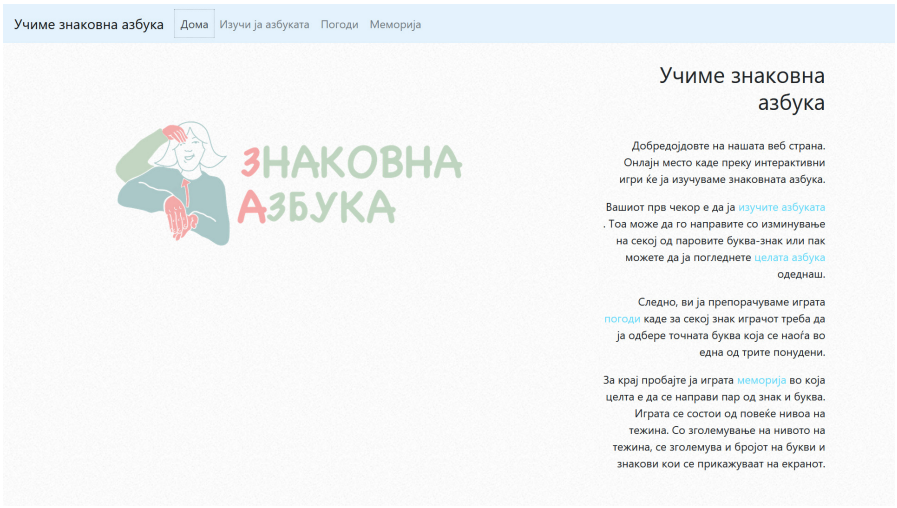


Figure 2. Web site and mobile application presenting part of the Macedonian sign language.

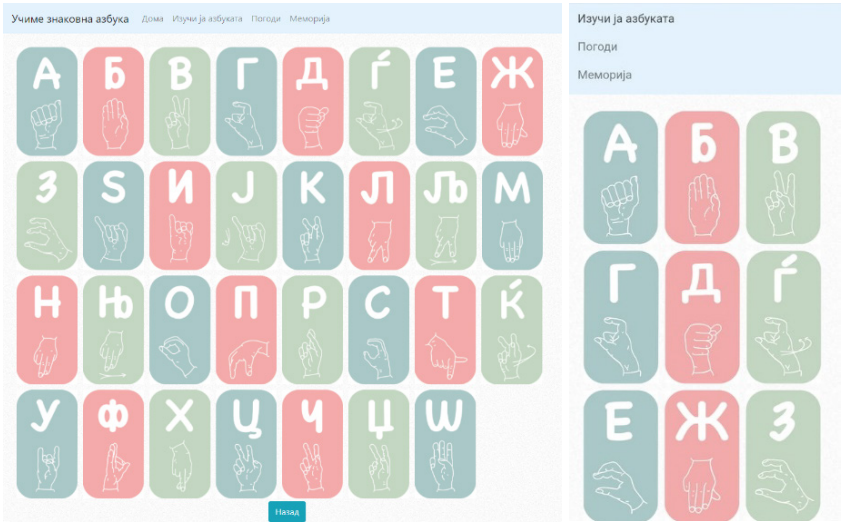
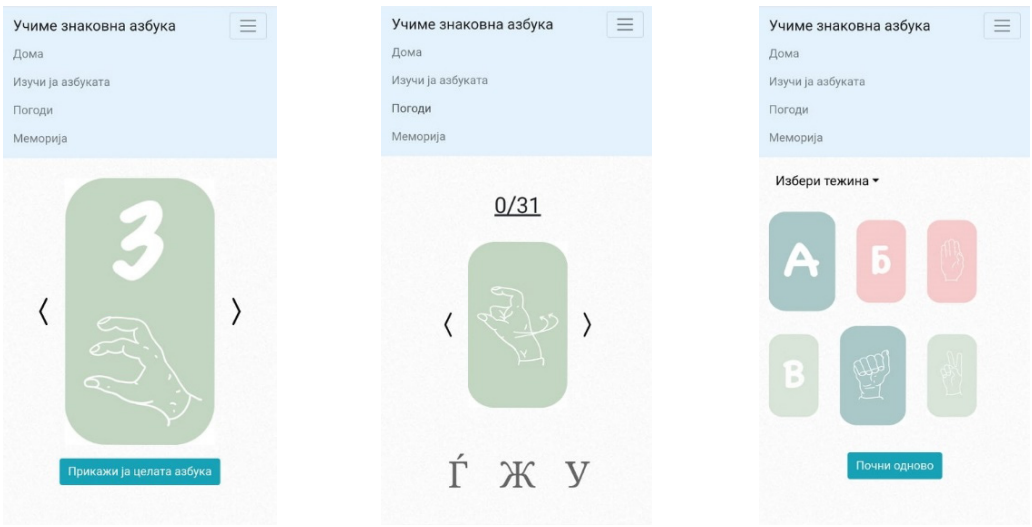


Figure 3. Presentation of both alphabets on one card, matching sign language character with the corresponding letter and memory game with one successfully matched pair.



In the next two games, the acquisition of frequent words and phrases will be enabled. Static and dynamic signs will be presented using 3D avatar visualization (Joksimoski, 2015). The assessment will initially be made by embedding the concept of memory games, based on matching an image with the corresponding sign. There are multiple approaches that can be utilized to achieve this. One approach is to use standard cameras and perform real-time analysis of the video, a field of active research. The second approach is to use specific sensors, like depth-based cameras (e.g., RealSense, Kinect).

Most of the children from day-care centre for children with DS in Skopje are speech unintelligible. They deserve equal rights to communicate and express their needs and feelings. The main prerequisites to assess the gamified approach are fulfilled. All the children with DS are eager to use multimodal communication, they are competent for gesture acquisition and production and they like playing mobile games. The day-care centre is currently closed due to Covid-19 pandemic. To start the evaluation of the application, the link has been sent to several children who are supposed to try it with assistance of their families. Except Niki, who is already familiar with the sign language alphabet, and who proudly presented his skills, other children had no interest to even try it. Therefore, the presentation is postponed for the reopening of the centre. Children feedback and suggestions by the language specialists will be crucial to make the improvements to current version and to offer it after the corrections and the enhancements on GooglePlay.

ACKNOWLEDGEMENT

This work is part of the project “Contribution to inclusive education of children with Down syndrome”, which is partially financed by the Faculty of Computer Science and Engineering (FSCE) at the Ss. Cyril and Methodius University. The authors thank Teodor Nikola Mladenovski, BSc at FSCE, who developed the application and Nina Pjevac, BArch at Manchester School of Architecture, who designed the images.

KEYWORDS: Down syndrome, educational software, sign language, verbal apraxia.

REFERENCES

Carl, M., et al. (2020). Vowel Acoustics and Speech Intelligibility in Young Adults with DS. *Journal of Speech, Language, and Hearing Research*, *63*(3), 674-687.

Chai, X., et al. (2013). SL recognition and translation with Kinect. In *IEEE Conf. on AFGR*, 655, 4.

de Almeida Barbosa, R., et al. (2018). Augmentative and alternative communication in children with DS: a systematic review. *BMC pediatrics*, *18*(1), 160.

Deckers, S., et al. (2017). Core vocabulary of young children with DS. *Augmentative and Alternative Communication*, *33*(2), 77-86.

Elsahar, Y., et al. (2019). Augmentative and alternative communication (AAC) advances: A review of configurations for individuals with a speech disability. *Sensors*, *19*(8), 1911.

Joksimoski, B., et al. (2015). Toward 3D Avatar Visualization of Macedonian SL. In *International Conference on ICT Innovations*, 195-203.

Kumin, L. (2006). Speech intelligibility and childhood verbal apraxia in children with DS. *DS research and practice*, *10*(1), 10.

Martin, G., et al. (2009). Language characteristics of individuals with DS. *Topics in language disorders*, *29*(2), 112.

Rice, M., Warren, S., & Betz, S. (2005). Language symptoms of developmental language disorders: An overview of autism, DS, fragile X, specific language impairment, and Williams syndrome. *Applied psycholinguistics*, *26*(1), 7-27.

Toth, A. (2009). Bridge of signs: Can sign language empower non-deaf children to triumph over their communication disabilities? *American annals of the deaf*, *154*(2), 85-95.

UNICEF. (1989). Convention on the Rights of the Child. Retrieved from https://ecommons.cornell.edu/bitstream/handle/1813/98856/crc.pdf?sequence=1

Zdravkova, K. (2020). Educational games for children with DS, In *Paradigm Shifts in ICT Ethics: Proceedings of the ETHICOMP* 2020*, 88-91. Universidad de La Rioja.

SOCIAL AFFORDANCES AND ETHICAL CHALLENGES IN MEDIATED COLLABORATIVE PLATFORMS

Peter Vistisen, Thessa Jensen

Aalborg University (Denmark)

vistisen@hum.aau.dk; thessa@hum.aau.dk

ABSTRACT

In this paper, we examine and analyse the affordances and social affordances of three online platforms, used as conference and learning tools: Zoom, Microsoft Teams, and Discord. With our grounding in the ontological ethics of Løgstrup and the theory of recognition by Honneth, our analysis suggests the need to focus less on the utility of available features of digital collaborative tools alone, but just as much on how the features encourage or inhibit desirable expressions among the participants. We propose this as a need to focus on community building as a main aspect for full online teaching and learning, as prerequisite before choosing and configuring didactic components. We analytically show how slight differences in the way the same feature is implemented in the three platforms can spark significantly different potentials for user interaction, expression and ultimate didactic participation. While all three examined online platforms provide feature-by-feature parity, only Discord shows clear social affordances encouraging multiple forms of expression and recognition, and thus enabling community feeling as being present *together, despite being apart.*

INTRODUCTION

In the wake of the COVID-19 pandemic, a wave of rapid digitalization swept over institutions all over the world. One of the ubiquitous changes is the adaption of digital collaboration tools to mediate meetings, teaching, collaborations, and social gatherings. At Aalborg University, Denmark, this happened in the middle of the Spring semester of 2020, leading to a focus on quickly mediating the curriculum, creating workspaces, and different possibilities for the interaction between teacher and learner. Less structured efforts were made in mediating a community of learners and providing a feeling of togetherness despite being apart. The issue of mediating culture and community through digital platforms is the focus of this paper. We argue this is a critical issue of both current and future situations of local lockdowns and reduced traveling due to upcoming sustainable agendas.

In this paper we examine and compare three online collaborative platforms regarding their sociability and functionality in the context of teaching and community building: Zoom, Microsoft Teams, and Discord. Of these, Zoom and Teams were implemented by Aalborg University during the first half of 2020 as the main teaching and learning applications for the transition to online coursework and project work. Because of the abrupt transition, these platforms were not fully integrated by university IT service at the time of the initial lockdown, causing us to use Discord as our first choice of platform at the time. Our main reasons were the students' knowledge of Discord, as they were already using it for their gaming and fandom communities, as well as its proven workability for online meetings, chatgroups, and collaborative groupwork, all of which are important in the context of problem-based learning (Kolmos, Fink & Krogh, 2007), the fundamental pedagogy at Aalborg University. With the reoccurring lockdowns, another aspect of the platforms became increasingly important: how could the online platforms help students and teachers deal with the isolation and stress of the ongoing

pandemic? Our focus shifted from creating an effective space of teaching and learning, to that of enabling a space for creating a community of learning, participation, and creation.

In the following, we will examine this focus through the optics of ethics and recognition theory, influenced strongly by Løgstrup's ontological ethics, and Honneth's structure of recognition. Løgstrup's ethical demand poses as a recognition of the Other through sovereign expressions of life in the form of trust, mercy, openness of speech, and sincerity (Løgstrup, 2013). The ethical challenge is found in this perspective, since each of the examined platforms has its uses in a learning environment, but the ethical demand of understanding and recognizing the Other, in our case students and fellow teachers, is not a given in the digital world. Løgstrup's ontological ethics with its focus on the dyadic face-to-face meeting and Honneth's theory of recognition, appear to be hampered by the mediation through screens and digital platforms (Løgstrup, 1997; Honneth, 2005). Of course, body language and facial mimic are missing or obscured, but it is possible to create communities in which its participants experience belonging and togetherness, despite the mediation through online platforms (Christensen & Jensen, 2018; Jensen, 2013). The need to be recognized as a person with needs and emotions, not just traits and abilities, may not be part of the curriculum as such. However, in the online environment of digital course and group work, it becomes a necessity. While the teacher concentrates on furthering the learner's abilities and knowledge, the learner, also, needs to be recognized as a person with the need of emotional support. Because of these demands, online platforms must ideally support and enable the building of communities in which primary relationships in the form of friendships, even love, can thrive side by side with a community of practice and solidarity. Honneth's modes of recognition, table 1, are showing these different dimensions of the personality, as well as the threats posed by a misrecognition of the same.

Table 1. Honneth's structure of relations of recognition.

Mode of recognition	**emotional support**	**cognitive respect**	**social esteem**
Dimension of personality	needs and emotions	moral responsibility	traits and abilities
Forms of recognition	primary relationships (love, friendship)	legal relations (rights)	community of value (solidarity)
Developmental potential	-	generalization, de-formalization	individualization, equalization
Practical relation-to-self	basic self-confidence	self-respect	self-esteem
Forms of disrespect	abuse and rape	denial of rights, exclusion	denigration, insult
Threatened component of personality	physical integrity	social integrity	honour, dignity

Source: Honneth (2005, p. 129).

The study into how the three collaborative platforms support the building of community and a milieu of recognition, is based on a participatory action research perspective (Chevalier & Buckles 2013). We have been teaching and 'hanging out' on the online platforms during the two semesters of 2020, and

the first semester of 2021, all which have been affected by the pandemic. The empirical basis of the study is done with inspiration to Hine's (2000) virtual ethnographic practice, and interviews with students and fellow teachers and researchers. Both authors have taught fully online courses during the three terms, as well as conducted project group guidance. While this limits the empirical basis for the paper's analysis being based on data akin to autoethnographic observations, we argue the combined 40+ years of teaching experience between the authors, as well as the colloquial engagement with the academic society throughout the data gathering process forms a professional and intersubjective grounding for the analytical findings. We were responsible for organizing courses and project work, as well as facilitating the creation and maintaining of a community for undergraduate and master's students. As shown by Haslam et al. (2021), these are some of the largest challenges during the lockdowns: enabling students to develop a sense of togetherness, community, and social closeness through online platforms to ensure their mental well-being.

Our research is centred on the functionality and the social affordances found in the three chosen platforms. How do these affordances enable the creation and maintaining of a community, which in turn enables learners to feel togetherness despite being apart?

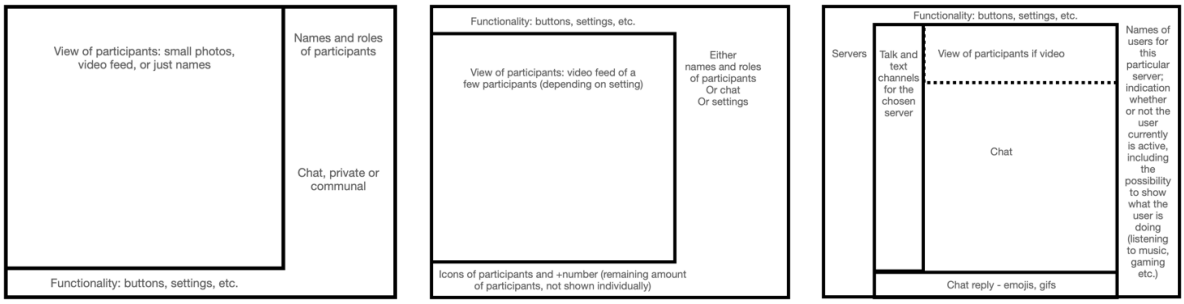
THE THREE PLATFORMS

The three platforms have feature parity in terms of technological functionalities, yet they differ on certain aspects. Zoom is a conference tool, Teams is a group work and teaching environment, while Discord is best known as a gamer and fandom communication platform. The main differences are to be found in the way the platforms afford different actions and cultures through miniscule, but important, differences in how the same feature set is implemented. A comparison of the three platforms with regard to online meetings is seen in figure 1. The figure shows a schematic of each platform when an online meeting is conducted. This is important, because without an actual video-meeting, Teams has major changes in its interface. When Teams is used as an archive or a chatroom only, it gives rise to major changes in interface, as shown in figure 2, which shows a comparison between Teams and Discord in the absence of a video-meeting. For Teams, these changes mean that both teachers and learners are frequently confused as could be witnessed during seminars, including a seminar for teachers on how to use.

Because of the different interface layers in Teams, the platform needs heavy facilitation in its initial use. The global changes depend on whether the user is in a certain Team, in a chat, a video call, editing a file, or using one of the many apps, which are integrated in the platform. Zoom's single-use orientation, video conference calls, makes the initial use easy to learn and understand. Discord provides a single interface with only minor changes depending on whether the user is in a video call or just using written and audio chat. As such, Discord's interface has a large amount of functionality present at all times, which provides a clear overview, easily learnt and adapted to the needs of group chats and collaborative work. While these differences could be waved off as non-consequential interface patterns that just have to be learned and remembered, there is a deeper level of how these slight differences can affect the didactic situation. If we consider for example the chat functionality through the lens of Buchanan's (2001) triad of aspects affecting the use qualities of a product: utility (is it useful?), usability (is it usable?) and desirability (does it retain my interest/create excitement?) we can see beyond the feature itself. In Zoom, the chat function is locked to the meeting itself, and only exists after enabling it via interaction with a button. It is only saved if the server allows for it and the user remembers to save it. In Teams, the chat has to be actively enabled by both the server and the users; it is automatically saved but is hidden within a container for the meeting after it ends. Discord, on the other hand, places the chat as the front and centre of its social engagement, layering

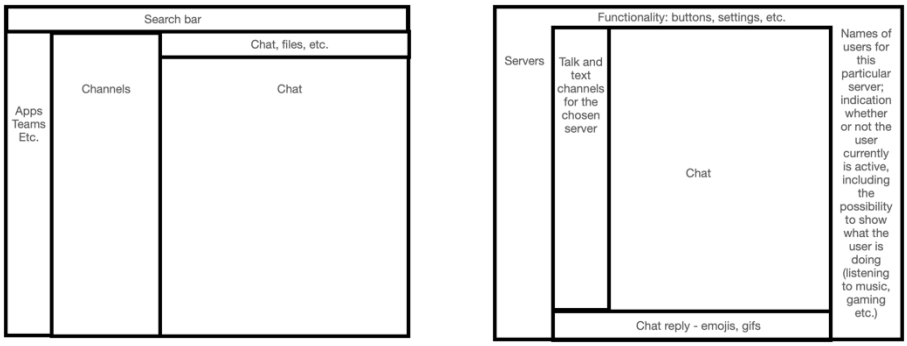
video and voice calls on top of the chat, which acts as a continuous stream within the channel itself. On the surface, the utility is the same of the three platforms, but the usability of how to access and retain the chat greatly affects the potential desirable role the chat function can play for the participants in the meeting – and if it can become a part of the didactic repertoire of the online collaboration

Figure 1. From left to right: Zoom, Teams, and Discord during a video-online meeting.



Source: own drawing based on the three platforms.

Figure 2. Teams and Discord without a video-meeting.



Source: own drawing based on the three platforms.

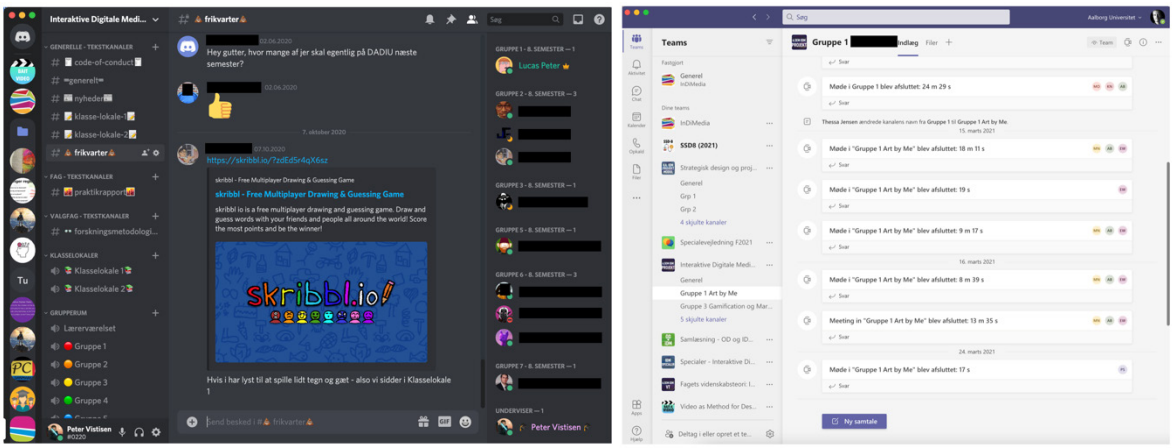
MURRAY'S AFFORDANCES

To examine further the different functionality and its influence on the creation and maintaining of a community, we now include Murray's (2012) grid of affordances. Murray distinguishes between four affordances in her grid: procedural, participatory, encyclopaedic, and spatial (Murray, 2012). She poses four questions to the analysis of the artifact: What does it do? What can I do? What are the boundaries of this domain? Where am I in relation to the whole? When comparing the three platforms in this regard, we can show how small features are implemented slightly different, but with rather significant implications for the user's interaction with both the software and with each other. Let us examine the chat functionality once again through this optic. While both Zoom and Microsoft Teams support chat functionality, Zoom allows participants to chat privately or choose to write out to everybody directly during the meeting. Microsoft Teams, on the other hand, allows for private chats as separate chat boxes, separate from the meeting, creating two virtual spaces existing simultaneously. In Discord, there is a mix of the two, where voice and text chats are visible to all, with full visibility of who are connected to what. Thus, a teacher or learner is able to see, who is present and interacting at any time.

What does it do, is Murray's procedural question. The three platforms have one similarity, offering video conference meetings. Besides that, they differ greatly in their procedural affordance. Discord's servers and channels offer an elaborate community-creating functionality, which enables group work and socializing, using transparency. At all times, everybody on a particular server can see who is present on Discord, and who is in certain text or voice channels. The text channels keep the conversations visible, enabling a feeling of presence, even when a user has been absent (Jensen, 2017). Teams' different layers offer different functionality, from a video call, to text chats, or archives of files for group and project work. Files can be edited in real time by the users of a particular team. However, the owner of the team has full control over what is possible to do for members of the team. Also, Teams offers an analysis on who has been present at what time, for how long. Data, which can be useful for teachers, may infringe the privacy of learners. Zoom's functionality is focussed solely on video-conference calls, with the possibility to create and work in break-out rooms, where smaller groups can discuss and work on given projects or tasks. Materials such as files or games need to be opened and worked on in their respective platforms or programmes. Teams and Zoom provide the possibility to record meetings, and Zoom's chat is, as mentioned above, lost when the meeting is closed.

What can I do? Murray's question regarding the participatory affordance is especially relevant in the context of community building. Zoom provides no possibility to create a lasting group work environment. Each video call must be separately recorded and stored outside the platform itself. There is no possibility to continue an ongoing discussion on the platform once a meeting is over. Teams provides the means to store material, files, videos; use other applications; and read and further discuss text chats. Both Zoom and Teams make a rudimentary reaction to chats or discussions possible through a limited amount of emojis and gifs–the latter only on Teams. In contrast, Discord allows for the use of a large amount of emojis and gifs, as well as giving users the option of being announced when entering a text channel. Users, teachers and learners alike, are able to write not only factual arguments, but express emotions, be funny, ironic, happy, or depressed. In our work as teachers, we could see a larger engagement by the students with each other on Discord than on the other two platforms. Especially in between courses and video-calls, Discord was the place where students did meet and hang out (figure 3).

Figure 3. Example of students using the didactically oriented Discord channel (left) to 'hang out' after class discussing both formal aspects of their studies, but also informal social interactions, such as quizzes etc. This is compared to the general tendency on Teams that channels are only active within the specific meetings, and not re-engage before the next meeting (right).

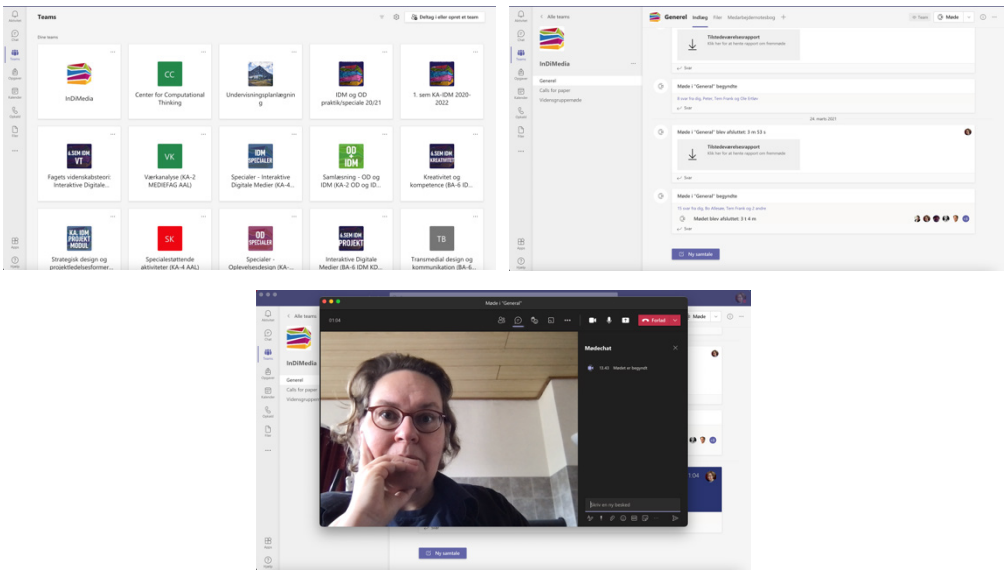


Source: authors' screen dumps.

What are the boundaries of this domain? This encyclopaedic question by Murray is easily answered for Zoom: its boundary is the actual ongoing video meeting. Once it has ended, the domain is shut down as well, leaving the participants to work through other platforms and programs. While other materials can be shared via screen share or by uploading a file to the chat (if this feature is chosen), no other way of communal participation in group work is possible. Discord's boundaries are defined by the platform as well: screen sharing or live streaming are possible, so is sharing documents and other materials, including links to videos, which are embedded into the ongoing chat. Discord provides a search function in any given chat, provided it has been indexed. Even so, the ongoing nature of the text chats is not feasible as a way of archiving, leaving students and teachers to use other platforms and programs to do so. Teams provides a filing system, which works as an archive. However, the files are listed in accordance with which file has last been accessed or uploaded, not alphabetical. Microsoft files and incorporated apps can be opened and edited directly in Teams, allowing for an extended domain. This shows an important aspect of the 'memory' of the virtual collaborative space where miniscule differences can greatly affect the social interactions. As an example, in Zoom you can only see the chat from the moment you join the meeting, leaving the user potentially puzzled if an unscheduled break was announced in the chat a minute before the user joining. The continuously saved chat on Discord can also present a potential usability issue for the didactic situation, since it can be hard for the participant to decode what part of the chat belongs to the specific activity and what belongs to a previous one – unless the teaching actively builds channels for each activity or uses the chat to meta-communicate.

Murray's last question, regarding the spatial affordance, is the most problematic for Teams: where am I in relation to the whole? Zoom has the video meeting and breakout rooms, providing an easy overview of where the user is at any given time. Discord shows the list of servers a user subscribes to, with the existing channels and text chat of the chosen server, at any time, including other users of the particular servers. Also, no matter which server is chosen, the basic interface of Discord does not change, even when a video meeting is called, the change in the interface is minimal. This is not the case with Teams. Its interface differs greatly depending on what part or app is currently in use (see figure 4). Furthermore, when in a team, a user needs to click on the Teams icon several times to return to an overall view of the teams the user subscribes to.

Figure 4. Some of Teams different interfaces.



Source: authors' screen dumps.

The three platforms show feature-by-feature similarities when they are analysed with the grid of media affordances, with their differences highlighted primarily in how the implementation increases or decreases usability, and less on the issues of what constitutes the potential ethical and recognition-oriented differences. We argue that these factors can be associated with the desirability differences among the platforms, and that these differences become obvious when examined through the emerging lens of social affordances.

SOCIAL AFFORDANCES - HOW A FEATURE IS MORE THAN THE SUM OF ITS PARTS

The differences among the platforms can be interpretated as the properties of the environment that act as socio-contextual facilitators relevant to the learner's interactions with the environments – what Krejns & Kirschner (2001) have labelled *social affordances*. When perceptible, social affordances invite learners to act in accordance with the perceived affordances i.e., to enter a communication episode and participate through the proper discursive premises.

Much of the use and design of online platforms for education depend on the teacher's pedagogical approaches. Davis and Chouinard present six views on how affordances afford especially for artifacts in social settings (Davis & Chouinard, 2016): requesting and demanding are bids, artifacts place on the subject; encouraging, discouraging, and refusing are ways, the artifacts respond to the desired action of a subject; allowing bids on both the subject, to act, and the artifact to respond to the desired action.

Zoom is mainly a conference tool, which should enable lectures and certain kinds of discussions among the participants; it is not meant as a tool for sociability. People can meet and network, but a community has to be in place before a Zoom meeting can enable a further development of a given community. Because Zoom lacks the option to save ongoing discussions and build an archive of any kind, it needs a work or learner culture to be firmly in place.

Microsoft Teams has the option to keep an archive, and participants are able to develop as a work group. Yet, few affordances are presented for building a community beyond the educational necessities. The platform is all about giving the teacher control of the learning environment, missing the opportunity to let learners develop a learning community on their own. Also, the platform converges different technologies like apps and programs into one place, which is very different from the approach of Zoom or Discord.

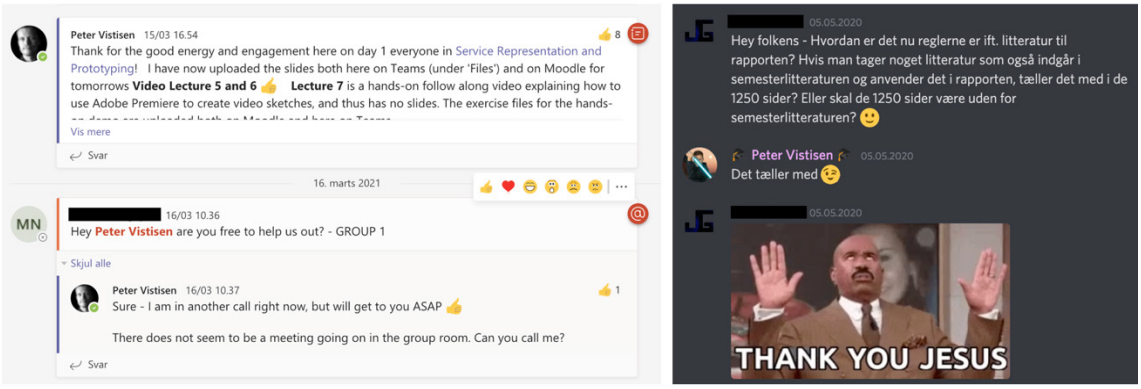
Discord offers a considerable higher level of sociability, mostly through the transparency of participation, which lets new participants decide where they would like to meet, and with whom. What is missing is the option to keep a searchable archive of files and information, something which necessarily would have to be kept on a different platform. Interestingly, this is how Discord is used by gamers and fans. The game or object of fandom is kept on gaming sites or fandom platforms, while the discussions and community building are done on Discord.

This highlights the intersubjective nature of social affordances, which Davis & Chouinard call a 'network of relations' that both enable (via features) and constrain (through discursive practices) technological capacities (Davis & Chouinard, 2016). Comparing Zoom, Microsoft Teams, Discord, and potentially other collaborative platforms through this lens reveals a web of relations among user perceptions, attitudes, and expectations – situating the differences between the platforms as relational processes among users, designers, environments, and the situations mediated. Discerning the platforms with regard to Honneth's structures of recognition (see table 1), all three platforms provide a recognition of cognitive respect, meaning the students and teachers have a moral responsibility towards each other, while legal relations and social integrity should be a given. Likewise, social esteem and the development of abilities and expressions of personal traits are possible on all

three platforms, enabling solidarity. However, emotional support and with it the development of friendships and recognition of needs and emotions, can be found mainly on Discord.

An intriguing example of relations regarding the social affordances are the differences in possible reactions of users towards each other's social presence. While video would be the obvious feature to provide feedback and interpersonal communication, students mostly avoid this option (as do teachers when in a seminar with fellow teachers or researchers). Instead, the written commentary in the chat is favoured when asking questions, and the ability to react with emoji 'thumbs up', 'smiley' etc. on all three platforms has become widespread. The difference is not in *what* is done feature-wise, but *how* it is performed as part of the social interactions. In Teams and Discord, the users can send animated gifs and memes through the chat. Teams is, by many IT-departments, restricted in what kind of gifs it allows to show. While this filtering can be applied to Discord servers, many channels mainly filter for violent or sexual content, and no restrictions are based on raunchy humour, political opinions etc, as is the case with some of the filtering applied on Teams. The topic of memes and online culture indicates that anti-social behaviour, trolling etc, are an unfortunate side-effect of unmoderated use of memes and other forms' visual expressions (Shifman, 2014; Phillips, 2016). We do argue the situation differs when considering the social affordances in an online didactic environment. In the first year of using Discord, we saw only one episode of anti-social behaviour communicated via memes. This was a spin-off between an already verbalised and written conflict between two students. The memes where not an escalation, but a continuation of a discourse which needed the teacher to step in, regardless of format. This is an outlier case, since the majority of meme use observed was indeed positive and provided an informal way for students to express small reactions of, for example, accept, encouragement, and appreciation, but also the occasional sarcastic reference to either the teachers or classmates. The important difference between the meme use on Discord and Teams was that Teams discourages the free use of memes and gifs as a mode of primary expression, instead promoting standardised emoji reactions, whereas Discord leans in with a social affordance of active encouragement.

Figure 5. The standardised emoji-reactions of Teams (left) providing a non-verbal mode of expression, but within a limited expressive space, as opposed to the active encouragement of free form visual expressions via memes on Discord (right). In the example, the student asks a formal question about the study regulations which the teacher answers, the student responding with a positive (but slightly sarcastic) meme.



Source: authors' screen dumps.

There is a point to be made in favour of social affordances. The simple emoji reactions of Teams leave little to be interpreted, and thus require significantly less literacy in online discourse, pop-culture, etc, while enabling the users to react through six well-established cultural symbols, still allowing for the use of memes and gifs in a more restricted manner. Discord's approach of both having an emoji-based reaction system (with 100+ emojis to choose from) and a less restricted use of memes gif reactions is putting responsibility onto both teacher and students when it comes to deciphering the cultural code of expression. In figure 5, the reaction meme could be interpreted purely sarcastically as an expression of too much or too little information given, or literal as viewing the teacher as the all-knowing presence in the channel. The middle ground, reading the *Jesus* comment as sarcastic and the *thank you* as a literal confirmation of the exact needed information, necessitates a literacy that covers not only usage of the specific meme, but the assigned roles and expectations among the users of the digital platform itself. The teacher needs to reflect on both the cultural literacy among the participating students, the media affordances of the platform itself, as well as the social affordances of how Teams and Discord respectively discourage and encourage a specific mode of expression to act and react appropriately with memes. But finding the right balance, it can be argued that allowing for, and actively engaging in, the students' use of memes and gifs is a potentially powerful way of creating a social sense of presence, by seeing memes, when they are used with honesty and respect, as Løgstrup's sovereign expressions of life within a shared literacy. As such, the differences between the platforms' allowance for handling the social affordances for memes and gifs might actually provide an explanation for the reluctance of users to turn on their web-cameras in live-sessions. This is an issue of *how* the platform promotes different participatory modalities for its users, and perhaps this short example indicates that some modes of expression may actually be more immediate, present, and appreciative than video? The example shows, how the binary utility notion of feature-by-feature parity is not enough to understand the differences between the platforms. Desirability of a particular platform is strongly influenced by medium-specific and social affordances, requiring a dedicated focus on cultural literacy and recognition to be used in an ethically and appreciative manner.

ETHICAL CHALLENGES IN ICT DESIGN AND USE – DISCUSSION OF AFFORDANCES

We point to an ethical challenge for the three platforms in their way of supporting an existing work and learning culture. Zoom and Teams both depend on existing communities, which need other platforms to conduct sociability and maintain their community culture. This other platform could be Discord, which is already used by fandom and gamer communities to support and develop communities. These communities create content, share their knowledge and experiences, as well as events on other platforms like Youtube, Twitter or Reddit, while their core community is found on Discord. In doing so we argue that the most crucial factor in the success of digital collaborative platforms is not 'what' the technology affords of user interactions, but 'how' the technology affords said interactions. In this 'how' we find the ethical demand of considering how the discursive space is formed by acknowledging the full complexity of the participants enacting their practice through the chosen digital platforms (Løgstrup, 1997; Vistisen & Jensen 2013).

Our analysis is based mainly on data gathered from students and teachers with a background in media and design research. Reservations have to be made regarding the need for an extensive media literacy, which includes knowledge and topicality of memes and emojis. Different groups of students and teachers will base their interactions on social affordances that, in turn, are based on the codes found in their social system of online culture. The need to embrace the digital tools of their online environment goes beyond merely joining a conference call if the participants of a learning community want to create and maintain an actual community of practice and participation.

Our main point is to put online communities before online learning. Creating a participatory and inclusive environment for recognition and trust, that should be *the* social affordance of any online platform used for learning. Our foundation for a digital pedagogy for digital teaching and learning platforms are the social affordances found in Løgstrup's ontological ethics and Honneth's theory of recognition.

We conclude by stating that the selection and implementation of digital collaborative platforms needs to take the mantra: *together, despite being apart* as a main objective in their design decisions. The social culture of a work or learner community needs to be transferred and transformed when work and teaching becomes digital.

KEYWORDS: social affordances, ethical design, online platforms, education, participatory culture.

REFERENCES

Buchanan, R. (2001). Design Research and the New Learning. Design Issues, 17(4), 3–23. https://doi.org/10.1162/07479360152681056

Davis, J. L., & Chouinard, J. B. (2016). Theorizing affordances: From request to refuse. *Bulletin of Science, Technology & Society*, 36(4), 241-248.

Chevalier, J.M. and Buckles, D.J. (2013) *Participatory Action Research: Theory and Methods for Engaged Inquiry*, Routledge.

Christensen, B. A., & Jensen, T. (2018). The JohnLock Conspiracy, fandom eschatology, and longing to belong. *Transformative Works and Cultures*, 27, [2]. https://doi.org/10.3983/twc.2018.1222

Haslam, C. R., Madsen, S., & Nielsen, J. A. (2021). Problem based learning during the COVID 19 pandemic. Can project groups save the day? *Communications of the Association for Information Systems*.

Hine, C. (2000). *Virtual ethnography*. SAGE.

Honneth, A. (2005). *The struggle for recognition: The moral grammar of social conflicts*. Cambridge: Polity Press.

Jensen, T. (2017). On the importance of presence within fandom spaces. *Journal of Fandom Studies*, 5(2), 141-156. [1]. https://doi.org/10.1386/jfs.5.2.141_1

Jensen, T. (2013). Designing for relationship: Fan fiction sites on the Internet. In H. Nykänen, O. P. Riis, & J. Zeller (Eds.), *Theoretical and Applied Ethics* (1 ed., Vol. 5, pp. 241-255). Aalborg Universitetsforlag. *Applied Philosophy / Anvendt Filosofi* 5(1). http://forlag.aau.dk/Shop/laering-og-uddannelse/theoretical-and-applied-ethics.aspx

Kolmos, A., & Fink, F. K. (2007). *The Aalborg PBL model: progress, diversity and challenges*. Aalborg: Aalborg University Press.

Krejns, K., & Kirschner, P. A. (2001). The social affordances of computer-supported collaborative learning environments—*IEEE Conference Publication*. 2001. https://ieeexplore.ieee.org/document/

Løgstrup, K. E. (1997). *The ethical demand*. Notre Dame: University of Notre Dame Press.

Løgstrup, K. E. (2013). *Opgør med Kierkegaard*. Aarhus: Klim.

Murray, J. H. (2012). *Inventing the medium: principles of interaction design as a cultural practice*. Mit Press.

Philipps, W. (2016). *This Is Why We Can't Have Nice Things. Mapping the Relationship between Online Trolling and Mainstream Culture*. Cambridge, Massachusetts: The MIT Press.

Shifman, L. (2014). *Memes in Digital Culture*. Cambridge, Massachusetts: MIT Press.

Vistisen, P., & Jensen, T. (2013). The Ethics of User Experience Design: Discussed by the Terms of Apathy, Sympathy, and Empathy. In A. Gerdes, T. W. Bynum, W. Fleishman, S. Rogerson, & G. Møldrup Nielsen (Eds.), *ETHICOMP 2013 Conference Proceedings*. Syddansk Universitetsforlag.

DEVELOPING AN EDUCATIONAL BRICK FOR DIGITAL ETHICS - A CASE STUDY-DRIVEN APPROACH

J. Paul Gibson, Yael Jacob, Damian Gordon, Dympna O'Sullivan

Institut Mines-Télécom (France), Institut Mines-Télécom (France),
Technological University of Dublin (Ireland), Technological University of Dublin (Ireland)

Paul.Gibson@Telecom-SudParis.eu; Yael.Jacob@Telecom-SudParis.eu;
Damian.X.Gordon@TUDublin.ie; Dympna.OSullivan@TUDublin.ie

ABSTRACT

In this paper we present the concept of re-usable educational bricks for the teaching of digital ethics. After describing the motivation behind the concept, we provide an overview of a standard template that can be used in the design of such a brick. We then briefly review the bricks that are at different stages of development, evaluation, and deployment, following this template. Finally, we conclude with a more detailed review of the development of a brick based on a case study which examines the use of "electronic pills" (e-pills) in the health industry. This case study falls within the computing topic of the Internet-of-things (IoT), and focuses on the ethical issues related to security and privacy.

INTRODUCTION

The work reported in this article is part of the Ethics4EU Erasmus+ Project (more details can be found at http://ethics4eu.eu) -

> The Ethics4EU Project is an Erasmus+ transnational project that will explore issues around teaching ethics in Computer Science. Ethics4EU will develop new curricula, best practices and learning resources for digital ethics for computer science students. It follows a 'train the trainer' model for up-skilling computer science lecturers across Europe.

The project objectives and deliverables are as follows: (i) a research report on *European values in Ethical technology*[1]; (ii) a research report on the *State of the Art of Teaching Ethics in Computer Science programmes*[2]; (iii) a comprehensive curriculum for teaching ethics in Computer Science; (iv) an open access online learning resources database of teaching and assessment strategies for teaching ethics in computer science; (v) an instructor guide to aid the delivery of material from the online resources database; and (vi) an online community of practice to facilitate discussion and experiences in delivering computer science ethics which will complement the online resource database and instructor guide.

The development of the educational bricks is based on the results from research reports (i) and (ii). It is the main contribution to objective (iii) and provides the material for the construction of the online teacher support platform, including a database (iv), instructor guide (v) and community of practice (vi).

[1] http://ethics4eu.eu/european-values-for-ethics-in-technology-research-report/

[2] http://ethics4eu.eu/outcomes/existing-competencies-in-the-teaching-of-ethics-in-computer-science-faculties-research-report/

The remainder of the paper is structured as follows. Section 2 provides the background and motivation for the work. Section 3 introduces the standard template for bricks which is based on a 4-dimensional classification method. Section 4 provides a general overview of the current state of brick development. Section 5 provides a more detailed description of one of these bricks – the *e-pills* case study. Section 6 concludes with a summary of the work that has been completed, and the work that has yet to be done.

MOTIVATION

The importance of well-integrating ethical aspects into computing programmes and modules/courses, as highlighted by Grosz et al. (2019) is well-established; and we are inspired by the research of Chuck Huff and C Dianne Martin (1995) which places emphasis on empathy, and students imagining the consequences of their own work and actions. Furthermore, we wish to encourage a more multi-disciplinary approach to teaching digital ethics as discussed in A.H. McGowan (2012). Our long-term goal is to provide a central repository (platform) of useful re-usable/adaptable education bricks for the teaching of digital ethics, following an "open" model – as proposed by Iiyoshi, Toru, and M. S. V. Kumar (2010) - such as seen with the creative commons approach. This platform will manage teaching material following good software engineering practices - as outlined in J. Paul Gibson, and Jean-Luc Raffy (2011) – for improved maintainability and sustainability.

In order to demonstrate the viability of such a platform we are currently developing a small set of six example bricks. The concept of an educational brick marries closely with that of learning objects, which Wiley (2000) defines as "small (relative to the size of an entire course) instructional components that can be reused a number of times in different learning contexts". It also fits well with the concept of 'distributed pedagogy' as used by Grosz et al. (2019). We hope that with this small set we can generate enough of a critical mass of academic users in order to build and maintain the repository.

EDUCATIONAL BRICKS FOR DIGITAL ETHICS – A STANDARD TEMPLATE

The need for a standard template is vitally important, particularly given the fact that the development of the bricks is being undertaken transnationally, with different bricks being drafted in different countries, and subsequently being reviewed and redrafted in other countries (including in the project partner organisation countries of France, Ireland, Italy, Sweden, and Switzerland).

The repository is currently under development, and the main requirements are for it to provide a rich set of features for searching for bricks, adding bricks, adapting/evolving different versions of bricks, composing bricks, etc. In figure 1, below, we see the brick that is found when we search our prototype system using **HCI** as a keyword: it is concerned with dark patterns in user interface design.

To help standardise and regularise the format and content of the bricks, the template includes two main sections: classification for searching purposes, and pedagogic issues for administrative purposes. Each brick can be associated with one or more case studies, and these are a key part of the classification. With respect to classification, we have four dimensions: ethical issues, academic domains, application domains and interdisciplinarity. Bricks, and case studies, may belong to multiple classes within each dimension.

We note that the template, and associated classifications, is not fixed. As we add more bricks, we expect the template to evolve as we validate its utility for meeting the requirements of the teacher support platform. We note that the platform prototype also permits users to provide feedback on bricks through a comment/chat functionality.

Figure 1. The ETHICS4EU Teacher Support Service Front-End.

Source: http://ethics4eu.eu/brick/dark-pattern-lesson/

Classification of Ethical Issues

The classification of ethical issues in the bricks can be done at two levels of granularity. Firstly, a high-level classification is based on the categories of interest identified in deliverable (ii) of the ETHICS4EU project: a) Origins of Digital Ethics, b) Digital Ethics Values, c) Data Ethics, d) AI Ethics, e) Ethics for Pervasive Computing, f) Ethics for Social Media, g) Governance and Legal Issues, h) Professional Ethics. Secondly, we provide a finer-grained classification based on the identification of ethics keywords in the case studies that we have incorporated in our teaching. This classification will expand and evolve as we add more educational bricks and studies. The current keyword list includes: AI super-intelligence; autonomy; bias, fairness, and transparency; discrimination; intellectual property; privacy and data protection; professionalism; safety and security; cyber-criminality and hacking; society, government, democracy, and environment.

This is the part of our classification scheme that requires more research, including the participation of a community of digital ethics teachers.

Classification of the Academic Domain

For the educational case studies, there is a requirement to match each study with educational requirements. The simplest way to do this is to list the knowledge areas (or skills) which would benefit from students interacting with the case study. The classification of knowledge areas is a complex task, and so we recommend using already developed taxonomies implicit in "bodies of knowledge" or "recommended curricula". The following list of four classification schemes are the most commonly used in Europe (and around the world), and should match with how most educational establishments classify educational content within the domain of "computing".

1. ACM Computing Curriculum[3]

[3] https://www.acm.org/education/curricula-recommendations

2. IEEE Curriculum (SWEBOK)[4]
3. European Research Council's Peer Evaluation (PE6) panel classifications of CS[5]
4. e-skills: The European Foundational ICT Body of Knowledge[6]

Classification of the Application Domain

Outside the academic domain, each brick case study must be classified under one or more application domains. As for the academic domain classification, we recommend using an existing scheme or standard. The three which we have found most useful are:

1. Global Industry Classification Standard (GICS)[7]
2. Industry Classification Benchmark (ICB)[8]
3. ISO Standards[9]

Classification of Interdisciplinarity

The ETHICS4EU approach encourages interdisciplinary teaching. For each case study brick it is very likely that academic disciplines other than computing could be involved in the teaching. For our initial set of brick developments, we have already identified potential for collaboration with the following academic disciplines - biology, physics, electronics, maths, psychology, history, law, medicine, philosophy, and engineering. Including an interdisciplinary classification explicitly acknowledges the opportunity for collaboration with other academic departments and colleagues.

Pedagogic Issues

With respect to pedagogic issues, we have five subsections, which correspond to the type of information that most high-level institutes record with respect to courses, modules and programs that they teach. A sixth is added to explicitly link to other bricks in our repository.

1. Academic Load
2. Pre-requisites
3. Learning Objectives (Ethical, Computing and Transverse)
4. Teaching and Evaluation Approach(es)
5. Support Material (For Teachers and Students)
6. Links to Other Bricks.

[4] https://www.computer.org/education/bodies-of-knowledge/software-engineering

[5] https://erc.europa.eu/sites/default/files/document/file/erc%20peer%20review%20evaluation%20panels.pdf

[6] http://ictprofessionalism.eu/wp-content/uploads/EU-Foundational-ICT-Body-of-Knowledge_Brochure_final.pdf

[7] https://www.spglobal.com/marketintelligence/en/documents/112727-gics-mapbook_2018_v3_letter_digitalspreads.pdf

[8] https://www.ftserussell.com/data/industry-classification-benchmark-icb

[9] https://www.iso.org/standards-catalogue/browse-by-ics.html

A SELECTION OF BRICKS – DEVELOPMENT, EVALUATION & DEPLOYMENT

As part of the ETHICS4EU project, the following bricks are at various stages of development and deployment.

1. *Foundations of Digital Ethics* is a brick which was developed as a pre-requisite to all other digital ethics bricks. Any student who has not had an introduction to ethics as part of their previous education experience will be required to follow the foundations brick. The brick has been developed, and validated by digital ethics experts, but it has not yet been deployed.
2. *Smart Pills* – this brick is detailed in section 5 of this paper. It has been developed and validated, and is due for deployment in May/June 2021.
3. *Software Certification, Accreditation and Testing* - professional ethics for software engineering. This brick is concerned with professional ethics in the domain of transport (aerospace and automobile) and is concerned with testing, certification and accreditation of software systems. The brick incorporates two main case studies – the Volkswagen emissions scandal, and the Boeing 747-Max crashes – which have attracted much media attention in recent years. These studies address the issue of the need for professionalism, and professional ethics, in the development of complex software. The brick is currently waiting for validation.
4. *Introduction to Programming - Algorithmic (AI) Bias* – is a brick for beginners to computing and computer programming. It illustrates that even the simplest algorithms can have bias (not just those based on complex AI and Machine Learning approaches). A central case-study is the use of algorithms for evaluating and assessing students during the COVID-19 pandemic. This brick has been developed, validated and deployed. Currently the teachers are analysing the feedback from the students.
5. *HCI-UX - Dark arts of interface design* – this brick collects a number of case studies from a range of different application domains to illustrate that dark patterns are ubiquitous. The brick examines the sometimes fuzzy boundary between unethical and illegal behaviour with regards to whether the use of such patterns should be considered a criminal activity. This brick has been developed, validated and deployed. Currently the teachers are analysing the feedback from the students.
6. *Autonomous Vehicles - more than just a trolley problem*. This brick is concerned with students' perceptions of autonomous cars, and whether the well-known trolley problem is a good way of teaching about the main ethical issues. The brick has been developed, validated and deployed. Initial analysis of the student feedback is very positive with respect to motivating students to be more aware of and concerned about digital ethics.

The next steps are to expand the community of bricks users. Two complementary approaches are being developed – (i) encourage re-use and evolution of existing bricks, and (ii) addition of new bricks. Already, there has been interest in – and initial development of – bricks which examine the following issues - Student Exam Surveillance and Proctoring, Facial Recognition, Public Surveillance, Tracking, Social Media and Fake news, Professionalism within Teaching and Research Ethics (conflicts of interest, publication practices), and Environmental Ethics - Cloud, AI, NFTs, Crypto-currencies and their impact on the planet.

A MORE DETAILED LOOK AT A BRICK – "E-PILLS"

One of the first bricks developed is one looking at "e- pills" (also known as "smart pills" or "robot pills" or "intelligent-pills"). These are a combination of a drug and a device, which can be described as "an oral tablet that incorporates some type of medical device, such as a microchip, that, for example, controls the release of the active pharmaceutical ingredient after ingestion" (Avery and Liu, 2011). This educational brick is aimed at 3rd/4th year engineering students who have chosen to specialise in information system management and development. As such, they participate in a module concerned with the architecture of complex systems, and apply their learning to developing a prototype system with a real industrial client, as part of a significant team project. In recent years, many of the team projects have incorporated technologies from the Internet-of-Things (IoT). Furthermore, the system requirements have become more and more demanding with respect to data protection and privacy (related to the GDPR in Europe). Finally, the students are becoming increasingly aware of the problem of such systems malfunctioning and the impact on the users.

As part of this module, the students are introduced to published research on general digital ethics issues - Ann Cavoukian et al. (2009), Gauthier Chassang (2017), Nancy Leveson (2020). They are also introduced to ethical issues through mainstream media reports on a wide range of technologies in different application domains. One of these studies is concerned with "smart pills" - Buffy Gorrilla (2017), Sandy Wash (2017). The students are then asked to research the main issues, and are provided with references to general papers on IoT and ethics - Ahmed AboBakr and Marianne A. Azer (2017), Josephina Antoniou and Andreas Andreou (2019) - and specific papers on medical ethical issues - Vinton G. Cerf. (2020), Kobi Leins et al. (2020), Brent Mittelstadt (2017), Julie Myers et al. (2008), Lily Hay Newman (2020), Ziad Obermeyer et al. (2019), Mark Stone (2019), and Daniel Wood et al. (2017).

Through discussion with teaching colleagues and students, there was general agreement that the "intelligent pills" provided an excellent case study with which to develop an educational brick on digital ethics. After playing around with various teaching ideas, the design of the brick was specified using the standard template, as follows. The student workload would be 9 hours contact time + 9 hours independent work. The pre-requisites are foundational knowledge of software engineering and networked/distributed system architectures. The computing learning objectives are: how to read documentation of IOT devices and evaluate whether there is coherency between natural language descriptions, formal technical specifications and the hardware. The ethical learning objectives are: consider who is responsible for the privacy of the sensor data; and the implications of the sensor being faulty/buggy. The transverse learning objectives are: communication skills and interaction with the media. The teaching domains are software engineering, architecture and IoT. The application domain is health. The ethical issues are security and privacy of data. The interdisciplinarity is with journalism and biology. The delivery mechanism/teaching approach is based upon students being involved in a debate with a journalist concerning whether the technical and ethical issues have been well-addressed in the general media. This will involve role-playing, following the advice from Diana Adela Martin et al. (2019). The evaluation is indirect – the students are evaluated through their project work, and one of the criteria is whether they have adequately considered the ethical issues. (The brick is currently being evaluated and refined, for first deployment at the end of the first semester of 2021.)

In figure 2, below, we see the header of the web site specific to the smart-pills brick. In this case the lecturer wished to provide their own front-end for access to the teaching material rather use the default interface provided by the platform. The brick is also included in the platform and will be found using any of the classification keywords; the platform then links to this autonomous web site. We chose not to force teachers to use the platform default template and encourage them to link their teaching material in whatever way is easiest for them. Currently this requires the platform

administration to classify the material by hand, but it is hoped to add automation to support this task. We also note that the web page uses scripts for navigation that are intended to link back to the platform. We hope to do this in the near future and provide a library of similar scripts for platform users.

Figure 2. The E-pills Web page

Source: http://jpaulgibson.synology.me/ETHICS4EU-Brick-SmartPills-TeacherWebSite/index.html

The readers are encouraged to visit the web site for this brick in order to see the different types of material that are provided to the teachers – teaching method support, scientific publications, books, journalistic articles from the popular press, social media posts, video and audio file links, etc. This material is being continually updated.

CONCLUSIONS

This paper has reported in the development of digital ethics educational bricks. This is work in progress, but initial results are very encouraging. We have reviewed the six initial bricks that have been developed, and provided more detail on the brick concerned with e-pills. Much more work is planned for the classification models – we are aware that the ethical classification is just an initial approach in order to quickly facilitate the construction of the teaching platform. We also acknowledge that we need to more formally specify the requirements of the platform in order to aid us in the construction of a community of digital ethics teacher resources. Finally, once the deployment of all initial six bricks has taken place, we intend to carry out an extensive evaluation and share the results.

ACKNOWLEDGEMENTS

The authors of this paper and the participants of the Ethics4EU project (http://ethics4eu.eu) gratefully acknowledge the support of the Erasmus+ programme of the European Union. The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

KEYWORDS: Digital Ethics Education, Case Studies, Internet-of-Things, Smart-Pills, Security, Privacy.

REFERENCES

AboBakr, Ahmed and Marianne A. Azer (2017). *IoT ethics challenges and legal issues*. International Conference on Computer Engineering and Systems (ICCES), pages 233–237. IEEE.

Antoniou, Josephina and Andreas Andreou (2019). Case study: The internet of things and ethics. *The Orbit Journal*, 2(2).

Cavoukian, Ann et al. (2009) *Privacy by design: The 7 foundational principles*. Information and Privacy Commissioner of Ontario, Canada, 5.

Avery, M. and Liu, D. (2011). Bringing smart pills to market: FDA regulation of ingestible drug/device combination products. *Food and Drug Law Journal*, 66(3): 329-352.

Cerf, Vinton G. (2020) *On the internet of medical things.* Communications of the ACM, 63(8):5, July 2020.

Chassang, Gauthier (2017). The impact of the EU general data protection regulation on scientific research. *ecancermedicalscience,* 11.

Gibson, J. Paul and Jean-Luc Raffy (2011). "A “Future-Proof” Postgraduate Software Engineering Programme: Maintainability Issues." *The Sixth International Conference on Software Engineering Advances* (ICSEA 11), Barcelona, Spain (October 2011).

Gorrilla, Buffy (2017) Gut feeling: the swallowable gut sensor that could replace a colonoscopy, *The Sydney Morning Herald.* January 20, 2017. Retrieved from https://web.archive.org/web/20201021122724/https://www.smh.com.au/national/gut-feeling-the-swallowable-gut-sensor-that-could-replace-a-colonoscopy-20170118-gttout.html

Grosz, B.J., Grant, D.G., Vredenburgh, K., Behrends, J., Hu, L., Simmons, A., & Waldo, J. (2019). Embedded EthiCS: integrating ethics across CS education. *Communications of the ACM*, 62(8): 54-61.

Huff, Chuck and C Dianne Martin (1995). Computing consequences: a framework for teaching ethical computing. *Communications of the ACM*, 38(12): 75-84.

Leins, Kobi, Chris Culnane, and Benjamin IP Rubinstein (2020). Tracking, tracing, trust: contem- plating mitigating the impact of covid-19 through technological interventions. *The Medical Journal of Australia*: p. 1.

Leveson, Nancy (2020). *Are you sure your software will not kill anyone*? *Communications of the ACM*, 63(2): 25–28.

Iiyoshi, Toru, and M. S. V. Kumar (2010). *Opening up education: The collective advancement of education through open technology, open content, and open knowledge*. The MIT Press.

Martin, Diana Adela, Eddie Conlon, and Brian Bowe (2019). The role of role-play in student aware- ness of the social dimension of the engineering profession. *European Journal of Engineering Education*, 44(6): 882-905.

McGowan, A.H. (2012). Teaching science and ethics to undergraduates: A multidisciplinary approach. *Science and Engineering Ethics*, pp. 1-9.

Mittelstadt, Brent (2017). Ethics of the health-related internet of things: a narrative review. *Journal of Ethics and Information Technology*, 19(3): 157–175.

Myers, Julie, Thomas R Frieden, Kamal M Bherwani, and Kelly J Henning (2008). Ethics in public health research: privacy and public health at risk: public health confidentiality in the digital age. *American Journal of public health*, 98(5):793–801.

Newman, Lily Hay (2020) Bluetooth-Related Flaws Threaten Dozens of Medical, *Wired*, 20 Feb. Retrieved from https://web.archive.org/web/20201021081330/https://securityintelligence.com/articles/the-potential-and-perils-of-the-iot-in-healthcare/

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464):447–453.

Stone, Mark (2019) The Potential and Perils of the IoT in Healthcare, *Security Intelligence*, November 21. Retrieved from https://securityintelligence.com/articles/the-potential-and-perils-of-the-iot-in-healthcare/

Wash, Sandy (2017) *FDA approves pill with sensor that digitally tracks if patients have ingested their medication*, US-FDA Press release, November 2017. Retrieved from https://www.fda.gov/news-events/press-announcements/fda-approves-pill-sensor-digitally-tracks-if-patients-have-ingested-their-medication

Wiley, D.A. (2000) Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. *The instructional use of learning objects*, 2830(435), pp. 1-35.

Wood, Daniel, Noah Apthorpe, and Nick Feamster (2017). *Cleartext data transmissions in consumer IoT medical devices*. In Proceedings of the 2017 Workshop on Internet of Things Security and Privacy, pp. 7-12.

2. Diversity and Inclusion in The New Normal

BRIDGERTON SERIES AS A PARADIGM OF FEMINIST CO-CREATION OF THE TELEVISION AUDIENCE

Graciela Padilla-Castillo, Aysel Zeynalova, Asunción Bernárdez-Rodal

Complutense University of Madrid (Spain)

gracielp@ucm.es; aysezey@ucm.es; asbernar@ccinf.ucm.es

INTRODUCTION

Bridgerton is a television series created by Chris Van Dusen, produced by Chris Van Dusen, Betsy Beers and Shonda Rhimes. Its first season premiered worldwide on 25 December 2020 on Netflix. The first season has eight episodes, as is the norm for all Netflix series, and a second season is already confirmed thanks to the fact that it racked up 82 million views in the first month of its premiere (Zorrilla, 2021). If it follows the structure of the collection of novels written by Julia Quinn, from which the original idea stems, the series could have up to 8 seasons, as there are 8 different books, each centred on the 8 Bridgerton children. The action takes place in London, around 1810, among the high society parties. It begins with the presentation of the young brides to the Queen of England and the appearance of a gossip bulletin, signed by an anonymous author: Lady Whistledown.

The plot, which at first sight does not seem very original, is adapted to our reality, offering original and necessary touches that form the basis of this investigation. On-screen, Shonda Rhimes innovates once again with a choral and multiracial cast, which she has previously imposed on all her series and which made her stand out especially from the first season of *Grey's Anatomy* onwards. Rhimes proposes blind casting so that actors and actresses are not chosen on the basis of physical characteristics imposed by the script. Thanks to this, the Queen of England, the Duke of Hastings (Sir Simon Basset), and his stepmother are black, breaking with history and with other fiction set in this era, always with Caucasian actors. Outside fiction, the promotion of the series has been original and humorous, on Twitter and Instagram, with the usual tone of the official Netflix account. However, viewers have created content based on the series, which has become more viral than Netflix's official publications. Specifically, the hashtag #bridgertonmusical, created by the user and singer @abigailbarlowww, has accumulated more than 185.7 million views to date, surpassing the audience figures for the series itself and standing as an excellent example of the importance of co-creation of content and the active role of the audience in cultural production.

With these precedents, this paper aims to study the Netflix series with three objectives: 1) to analyse the original ideas offered by Bridgerton in terms of its female and multiracial cast; 2) to compare this proposal with the series previously produced by Shonda Rhimes and look for the proposals of female empowerment that she always bets on; and 3) to study the cultural produsage or co-creation of content of the audience of the series in TikTok, commenting on the typologies of proposals and the messages of female empowerment that they promulgate. These three objectives will be addressed with a qualitative methodology in three phases: 1) longitudinal study of the concepts of cultural produsage, audience co-creation and empowerment; 2) commentary on the figure and filmography of Shonda Rhimes; and 3) exploration of the most used hashtags of the series on TikTok, listing the proposals, formats, tone and scope.

ON AUDIENCE CO-CREATION AND THE ACTIVE ROLE OF AUDIENCES

Taking advantage of the technologies and innovative tools provided by social networks, consumers have become protagonists and content generators. In 1979, Alvin Toffler, in his manuscript *The Third Wave*, argued that consumers are a phenomenon of the industrial era. He maintained that in the post-industrial era, consumer behaviour was moving towards prosumers, those who, in addition to consuming, produced for themselves. More than forty years later, this concept is still valid and more plausible than ever.

Verwey (2015) highlights the collaborative and expressive nature of interactive media and technology, and the ability they offer users to participate in the production and publication of branded content within digital affinity communities. He adds that this landscape presents greater possibilities for self-expression, as well as unlimited opportunities for participation in determining and influencing the narratives being developed at any given moment. And Marinas (2019) notes that with social networks, new forms of communication and learning have emerged that were previously unknown.

Hatch and Schultz (2010) were pioneers in defining the 4 basic pillars of the co-creation process: dialogue, access to information, transparency and risk. They propose a simple model of co-creation in two dimensions: engagement between the company and its stakeholders and information provided by the company. They conclude that there is a growing interest on the part of companies in offering multiple channels that allow them to create a commitment or link between the company and its stakeholders; generating dialogue with their publics through the channels through which the company is accessed. On the other hand, they confirm that stakeholders are increasingly demanding more information about organisations and brands, even with the risk that this may entail for the corporate image.

In a similar vein, Ramaswamy and Ozcan (2016) concluded that in the traditional brand value creation process, companies viewed their audiences as passive recipients. However, in the latest brand co-creation processes, all stakeholders play a more active role. They contribute their opinions to the creation of brand value together with the company. The authors recommend that company managers set up brand experiences where individuals can carry out co-creation actions to increase brand value. The same authors, in another paper from the same year (Ramaswamy and Ozcan, 2016) focused on developing the concept of "joint experience of agents in the creation of brand value". They delved into how to involve different stakeholders, personally and collectively, in brand value creation, expanding the way in which the firm connects brand value creation opportunities with brand resources.

In this context, digital engagement platforms are fundamental, specifically designed as a system of people, elements, interfaces and processes that favour the development of interactive environments to intensify the joint experience and action of co-creators and generate mutually valuable results for all participants and agents in the brand value co-creation system. It is interesting how they insist on the difference between "agent" and "actor", understanding that the "agent" has the capacity to act motivated by its engagement as an individual who reproduces and transforms its structural environment through its relationships with the same environment and with the other agents interacting in that environment.

Hsieh and Chang (2016), meanwhile, integrated perceived psychological benefits and distinctive motivations into consumers' brand co-creation process from self-determination theory (Deci and Ryan, 1980) and implicit self-esteem theory (Greenwald and Bnaji, 1995). They found that: (1) high self-connectedness to the brand facilitates brand co-creation engagement; (2) both autonomy and perceived personal competence or aptitude in brand co-creation tasks are positively associated with brand co-creation engagement; and (3) brand co-creation tasks that bring a perception of relatedness or affinity among co-creation team members also facilitate brand co-creation engagement to be

established, which, in turn, increases purchase intention and other positive attitudes toward the brand.

Erdem et al. (2016) studied the control companies have over their own brands; the new relationships they are establishing with consumers; the risks of co-branding; and the threats of this new process for brand management. They consider it essential to study whether co-creation affects the growth of brands and whether brand ownership is diluted. They also address the question of how companies should develop integrated communications strategies to better reflect the wide variety of digital options. Tajvidi, Wang, Hajli and Love (2017) also propose a model of brand co-creation in which consumers' relationship with each other and with brands positively affects the sense of belonging to a community and facilitates brand co-creation in electronic environments. Following the precedent of Prahalad and Ramaswamy (2004), they assume that co-creation of brand equity is deeply rooted in the concept of co-creation of value. They take up Prahalad and Ramaswamy (2004) definition of co-creation as the collaboration between a customer and a supplier in the activities of co-design, co-design and co-development of new products. They point out that the academic literature traditionally recognises that value can be created in the co-creation process, when customers move from being a passive audience to being a social partner in the co-creation process, and that value can be created in the co-creation process when customers move from being a passive audience to a social partner in the co-creation process. In this way, they accept that value creation between customers and suppliers is based on a unique experiential environment in which customers engage in dialogue and interaction with their suppliers, as well as access to their resources (Prahalad and Ramaswamy, 2004).

ON ACTIVE SOCIAL MEDIA AUDIENCES: TIKTOK

The social, economic and cultural context almost a year and a half after the World Health Organization declared COVID-19 a pandemic respiratory disease is on many levels very different. The course of events forced half of the world's population, 3.9 billion people, to respect some form of confinement (France24, 2020). In Spain, the government declared a state of alarm, limiting, among other things, social relations and the free movement of the population (Government of Spain 2020).

The *23rd Surfers on the Net Report* (in original Spanish, *23º Informe Navegantes en la Red*), conducted by the Association for Media Research (in original Spanish, AIMC, 2021), details the growth of Internet connectivity with a constant frequency, according to 47.48% of respondents, and 92.30% who connect several times a day. The smartphone is the most chosen access device, according to 93.1% of respondents, and the laptop (71.4%) or desktop (50.75%) occupy, respectively, the second and third place.

Bearing in mind that 87% of internet users aged between 16 and 65 use social networks (IAB Spain, 2021), we can deduce that a large part of this connection time is spent on these platforms, which were born to socialise, but which have developed and evolved towards marketing and interactivity at all times. On this point, in the *Social Networking Study 2021* (in original Spanish, *Estudio de Redes Sociales 2021*, IAB Spain, 2021) there is a categorisation that indicates the activities most carried out by users. 81% seek to be entertained, 77% seek to interact and 66% seek information.

Consequently, social networks have adapted and evolved to positively address all these needs and this is demonstrated by the growth of users and the increase in connection time mentioned in the first lines. Also in the Spanish case, reports by Hootsuite and We Are Social, The Social Media Family, IAB Spain, Statista or Inesdi highlighted TikTok as the social network with the highest growth. In fact, the app has been named as the 'Social Network Revolution' of 2020 and has increased all its metrics by 3 or 4 compared to 2019 (Weimann and Masri, 2020).

This app (still under the name Douyin in the People's Republic of China) is part of the ByteDance Company. The Asian corporation has its tax domicile in the Cayman Islands, although its headquarters are in Beijing. To date, it has shareholders from all continents, through the venture capital funds that have invested in the company. Among the most prominent are Sequoia Capital, Kohlberg Kravis Roberts and General Atlantic. Its operations are divided into offices in different countries, although its main revenue base is China (Liu and Yu, 2020). Projections for this year 2021 announce spectacular revenues, taking into account the rise to $35 billion in 2020 (Huang, 2021).

Moreover, the keys to TikTok are numerous, but simple. Detecting and understanding them makes its growth and virality understandable (Padilla-Castillo, 2021):

- No need to have an account or register. Anyone, by downloading the app, can watch videos, download them, forward them, press likes or report them.
- It has videos on every imaginable topic.
- The algorithm suggests videos "For you" and has memory: it not only offers you topics of videos you have watched in the last hours, but also in the previous weeks or months, in case you had forgotten them.
- A video can go viral without the user who created it being an influencer or having thousands of followers.
- The humorous and cathartic component has been fundamental, as a pastime during sanitary confinement.
- It is a very simple network to use and does not offer navigation tutorials, promoting precisely its simplicity.
- The videos are very short, up to one minute long, and can be viewed at any time and occasion of the day.
- It offers simple and viral challenges, seeking the co-creation of the public, so that they imitate them following the original idea.
- It does not demonize plagiarism, but virality: you can make a video with the background, audio and music of another user, or you can make a duet with him/her with the shared screen.
- Video editing is very simple, with numerous filters, always free and completely openly available. They are on the platform itself and the user does not require other image, video or audio editing programs. No previous technical knowledge is required either.

In these circumstances, users quickly feel like protagonists and users. They see that they are part of the series, that they can influence its continuity and its history. For some, TikTok allows a direct and simple form of monetization, an artistic space that also becomes a professional space.

ABOUT THE LIFE AND WORK OF SHONDA RHIMES

Shonda Rhimes was born in Chicago in 1970 to a family of five siblings and parents in academia: her mother was a professor and her father a manager at the University of Chicago. She graduated from Dartmouth College and won a scholarship to the University of Southern California for a master's degree in film and television screenwriting. After graduating, he was unemployed for several months and had only very short-term jobs. The exception was the documentary *Hank Aaron: Chasing the Dream* (1995)

and the short film *Blossoms and Veils* (1998), acquired by New Line Cinema. Soon after, he got his break as a screenwriter for the 1999 TV movie *Dorothy Dandridge*, which made actress Halle Berry famous. And from there to teen films with two titles: *Crossroads: To the End* (starring Britney Spears) in 2002, and *The Princess Diaries 2: Royal Engagement* in 2004. It did not find its place and the titles did not perform at the box office as expected.

Until the first episode of *Grey's Anatomy* aired on 27 March 2005, and her CV changed forever. The series, meant to be a mid-season break or transition product, won over audiences and quickly gained its own space. It moved from Sunday to Thursday, a prime time slot in the American prime time. It revived the medical series, along with *House, M.D.*, which had not enjoyed such splendour since *ER*. It made Shonda Rhimes one of the 100 most influential people in the world, according to *Time* magazine's ranking. All thanks to a choral and multiracial series, with the young and inexperienced doctors learning about the profession and about human and love relationships.

The success of the series would lead to the appearance of a sequel or spin-off. Dr. Addison Montgomery left *Grey's Anatomy* for her own show, *Private Practice*. In fiction, she moved from Seattle to Los Angeles. In reality, the then husband of the lead actress, an ABC executive, proposed to Shonda Rhimes that she create a whole new series, as *Grey's Anatomy* had done. It was made to show off actress Kate Walsh and Rhimes was happy because she was also producing it with her company, ShondaLand. Finally, in 2013, and after 6 seasons, the series was cancelled with 11 international television awards and more than a score of nominations.

It was not to be Rhimes' great success as the predecessor series, *Grey's Anatomy*, remains her benchmark and her other successes have not been able to surpass it. At the close of this paper, the series has 381 episodes, 18 seasons, 16 years on the air, 4 Emmy Awards, 76 other international television awards and more than 230 nominations. Its protagonist, Ellen Pompeo, who gives life to the Grey of the title, is also a producer of the title and one of the most influential women in the United States. In fiction, she survived the death of her first partner and father of her children; and in media interviews, she has often referred to the rumours that guessed, without reason, that she alone could not carry the weight of the entire series without a male partner.

In addition to this series, and before *Bridgerton*, Shonda Rhimes has triumphed with many other titles that endorse her good eye for choosing and implementing television fiction projects. After *Private Practice* came *The Catch*, which lasted only two seasons (2016 and 2017), but gave great international popularity to its protagonist, actress Mireille Enos. Known for her brilliant role as a police inspector in the dramatic and black series *The Killing* (2011-2014), the decision to cast her as the protagonist was by no means gratuitous. In fact, after *The Catch* she has moved on to another powerful role, in *Hanna* (2019-2021).

In parallel, Shonda Rhimes developed *Scandal*, another feminist and multiracial milestone in her filmography. It was premiered by ABC (American Broadcasting Company), on April 5, 2012, and remained on air for 7 seasons, until 2018. It is a drama series, with episodes of 43 minutes each. Each installment contains a self-conclusive plot: the protagonist, Olivia Pope, and her team must fix a communication crisis of a client. This crisis opens each episode and is resolved at the end of the episode. Apart from these episodic plots, with episodic characters, there are long plots for each season, which affect the main characters and the empathy they can and do provoke in the viewers.

Scandal was filmed in Los Angeles (California), although the fiction was developed in Washington, D.C., always around the White House. It was produced by ShondaLand, Shonda Rhimes' production company, and ABC Studios, as was the case with *Grey's Anatomy* and *Private Practice*. In addition to having the audience's favor, during those 7 seasons, the series deserved 2 Emmy Awards, 33 other

television awards and more than 70 nominations. Almost all of them were to recognize the role of black actress Kerry Washington, playing the protagonist, Olivia Pope.

Pope is the alter ego of Judy Smith, Rhimes' other half for this project. She was born in 1958, in Washington, D.C. She studied Public Relations at Boston University and after graduating, she was a communications assistant at the College of Obstetricians and Gynecologists in her hometown. Shortly thereafter, in 1989, she became a deputy spokesperson for the U.S. Attorney for the District of Columbia. Interestingly, this position would have great importance, years later, in the series Scandal, where the same character appears: David Rosen played by actor Joshua Malina.

She was in that position for two years because in 1991, she became part of the cabinet of the President of the United States, George H.W. Bush. The country's forty-first president served from 1989 to 1993. Previously, he had served as Ronald Reagan's vice president from 1981 to 1989 and as director of the Central Intelligence Agency (CIA) from 1975 to 1977. It is no coincidence that the CIA is also an essential part of the plot of *Scandal*.

In the Bush administration, international policy was a fundamental part. The President decided to invade Panama on December 20, 1989, in *Operation Just Cause*, to capture General Manuel Antonio Noriega, military leader and then dictator of the country. He also initiated the Gulf War in the summer of 1990, with a coalition of 31 countries and the authorization of the United Nations, in response to the invasion of Kuwait by the then Iraqi leader Saddam Hussein. The conflict lasted until February 1991. And in June 1993, the president ordered the bombing of Iraq again, in retaliation against the alleged plot to end his life. Yet again, reality would invade fiction because President Fitzgerald Grant, in *Scandal*, played by Tony Goldwyn, lives the same threat.

The real conflicts with Iraq would return in 1998, during Bill Clinton's administration and *Operation Desert Fox*; and during the administration of George H.W. Bush's son, George W. Bush, in 1998. However, those crises were far removed from the work of Judy Smith, who was closely involved in the first Gulf War. After finishing her term and leaving the White House, she founded her own firm: *Smith & Company*, specializing in crisis management and public relations. She would never leave politics, as she advised high-ranking officials and Monica Lewinsky, who declared having nine sexual encounters with then President Bill Clinton, between 1995 and 1997.

Smith was also vice-president of the communications office of the American television network NBC (National Broadcasting Company). Thanks to this job and her television contacts, in 2009, she was introduced to Shonda Rhimes and her partner, Betsy Beers. The meeting was scheduled to last about half an hour. But the three women talked for several hours and the germ of Scandal was born. Smith would become executive producer of the series, along with Rhimes, and of course, plot and script consultant.

Rhimes tried her hand at historical TV drama with *Still Star-Crossed*, in 2017, but it only ran for one season. And after *Scandal* ended, she got her start in legal drama with 2-season *For the People* in 2018-2019. These series became unremarkable flops because as the end of *Scandal* approached, Rhimes returned to hit the bullseye with *How to Get Away with Murder*. The series has had 6 seasons, through 2020, and its lead is another strong, empowered, black woman. Viola Davis plays criminal defence attorney Annalise Keating for 90 episodes and has earned the series an Emmy Award, 17 other international awards, and 77 other nominations at the close of this project.

Finally, before *Bridgerton*, Rhimes returned to drama-sanitary series with *Station 19*, set in a fire station in Seattle, Washington. This proposal stands out for giving a broad protagonism to women in a profession that is usually represented in fiction, almost always, by men. It has been on the air for 5

seasons, until 2021, and the parity and multiracial cast includes actors of black and Latino descent, which is another important plea for equality and the reduction of discrimination by gender or race.

BRIDGERTON AT TIKTOK: PARADIGM OF AUDIENCE CO-CREATION

According to the objectives of the paper, after the longitudinal study of the concepts of cultural produsage, audience co-creation and empowerment and the commentary on the figure and filmography of Shonda Rhimes, we offer an exploration of the most used hashtags of the *Bridgerton* series on TikTok, listing the proposals, formats, tone and scope.

In order to be as aseptic as possible, this search has been done from a mobile phone with the app downloaded, but without having a registered user account. By Popularity, TikTok offers 10 videos for #bridgerton, as the most recommended, without logging in or having any knowledge of the user's tastes and habits:

1. “Bridgerton in real life” (@annahosp): 2.5M likes, 55.1K comments and 76.9K shared.
2. “bridgerton” (@gloriday): 600.6K likes, 7,865 comments and 32,3K shares.
3. “Me watching the First episode of Bridgerton” (@rozyqueenofcups): 2.2M likes, 22.4K comments and 126.9K shares.
4. “Just girly things” (@keirayasmin6): 23.5K likes, 438 comments and 791 shares.
5. “Currently facing Bridgerton” (@hoesloveashanti): 253.6K likes, 2,892 comments and 3,363 shares.
6. “Your grace!” (@mishdontkillmyvibe): 382.6K likes, 5,079 comments and 13.0K shares.
7. “Not me rewatching it the 3rd time” (@futuremilfwithoutkids): 1.4M likes, 23.7K comments and 37.5K shares.
8. “Duque” (@s.netiflix): 44.9K likes, 704 comments and 5,842 shares.
9. “Simon & Daphne” (@bridgertonlatino): 13.3K likes, 65 comments and 578 shares.
10. “It’s on Netflix” (@ubiiinetflix): 200.5K likes, 812 comments and 991 shares.

For Users, we list the 10 accounts offered by TikTok, in the same decreasing order in which they appear in the app for the #bridgerton search:

1. @bridgertonlatino: 39.7Kfollowers and 44 videos.
2. @bridgerton.romania: 4,803 followers and 17 videos.
3. @iiburnforyou: 1,699 followers and 71 videos.
4. @osbridgertons.2021: 2,112 followers and 9 videos.
5. @bridgerton_ofc: 6,814 followers and 56 videos.
6. @ bridgerton20: 13,7K followers and 99 videos.
7. @ bridgertons.0: 5,963 followers and 79 videos.
8. @bridgerton_life: 908 followers and 38 videos.
9. @ netflixbridgerton: 130 followers and 6 videos.
10. @bridgertonlovers: 344 followers and 1 video.

It is very striking, as in *Popularity*, that these accounts do not appear in order of number of followers, number of videos or number of likes. The social network's secret algorithm offers them in the same order, for the #bridgerton search, understanding geographical and language issues, as there is no previous history of searches or views on the account used for the study. As proposed for Popularity, it would be interesting to delve deeper into each of these accounts, in future research, and to understand why TikTok presents them in this decreasing order. Above all, it is striking that there are accounts with few followers, with few videos or a single video, or with few followers and dozens of videos at the same time. It would also be very interesting to draw up a ratio of followers/number of videos and to analyse in depth whether the series, within an account, is dealt with in one or several videos; and whether it is related to other series, feminist issues or other topics.

Specifically, the hashtag #bridgertonmusical, created by the user and singer @abigailbarlowww, has accumulated more than 185.7 million views to date, surpassing the audience figures for the series itself and standing as an excellent example of the importance of co-creation of content and the active role of the audience in cultural production. Its creator defines herself in these terms: “Songwriter girl. Co-creator of #bridgertonmusical. Check out our website below! www.barlowandbear.com” (@abigailbarlowww, 2021). She has 2.3M followers, 41.5M likes and several hundred videos: her own and duets with her followers.

The key to her success is to make songs of musical genre, sung and spoken, mixing her own text and lines from the script of the series. She appears in clothes of our present time, not like the protagonists of the series, and her hair is dyed purple. She always sings live, without editing, and with the accompaniment of an instrument (almost always piano) in some cases. She alone plays several characters when in her videos she refers to a dialogue between the characters in the series.

The partner playing the piano in some of the videos is Emily Bear, her partner in the company Barlow & Bear. On their website, they explain that they have broken the glass ceiling for women by offering a Broadway-style musical on a social network, free of charge, and with many more viewers than a real show gets (Barlow & Bear, 2021).

En su página web también explican, en unas líneas, el origen de su éxito: #bridgertonmusical. Abigail Barlow supo que era una historia perfecta para el escenario teatral-musical y escribió los temas "Oceans Away" y "I Burn For You”, que cantó en TikTok. “Absolutamente asombrada por la respuesta” (Barlow & Bear, 2021), contacto con Bear para desarrollar más temas a partir de las tramas de la primera temporada. Respondieron usuarios anónimos, cantantes, Netflix, la autora de los libros (Julia Quinn) y los creadores de la música original de la serie, Pasek and Paul, que la han felicitado públicamente por crear un nuevo género musical.

CONCLUSIONS: TIKTOK, CO-CREATION AND THE ROLE OF THE AUDIENCE

This research had three objectives: 1) to analyse the original ideas offered by Bridgerton in terms of its female and multiracial cast; 2) to compare this proposal with the series previously produced by Shonda Rhimes and look for the proposals of female empowerment that she always bets on; and 3) to study the cultural produsage or co-creation of content of the audience of the series in TikTok, commenting on the typologies of proposals and the messages of female empowerment that they promulgate. According to the methodology and the results discussed in the previous lines, the three objectives have been met, although the research has revealed new and very interesting prospects for further research. On TikTok we have discovered that its algorithm can make any video, any hashtag and any account go viral, regardless of these three aspects of each other. When a user goes viral, like @abigailbarlowww, it is understandable that her videos also go viral, but there is no such reciprocity all the time, and she

herself achieved worldwide fame thanks to just two videos. As for Shonda Rhimes, after analysing the role of women and race in her series, we see courage and hopefully proposals that many series may follow. *Bridgerton* is her latest success, but it is not her first, and it is striking that people criticise the parity and ethnicity of her cast when it should be a trend to follow in the times we live in.

The union of the social network, the series and its cast, and the original co-creation of the audience give rise to a parallel world to the Netflix screen, where the series takes on new life, shaped by its users, and becomes bigger and multiplies. Social networks, like TikTok in this case, offer a dynamic, changing, original life, with many approaches and readings, where the audience becomes the creator. New stories are constructed that may not be incorporated into the series, but that enjoy thousands or millions of views. And in the midst of this virality, feminist empowerment becomes greater, as Rhimes has been trying to do since her first work. Her series break the gender gap and the social dialogue of her series contribute to that ceiling not only being broken in fiction, but also in the reality of the audience. @abigailbarlowww is the best example because she has created a musical on the Chinese social network that no one would have produced for her on Broadway. Now, she makes a living from the music she dreamed of writing and singing. If only many more series would allow for such success stories and paradigm shifts, which show the best consequences of television, so unfairly demonised and undervalued.

ACKNOWLEDGEMENT

This research has been carried out as part of the competitive research project: “Produsage cultural en las redes sociales: industria, consumo popular y alfabetización audiovisual de la juventud española”. R&D project of the Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia 2018-2022. Reference: FEM2017-83302-C3-3-P. Ministry of Economy and Competitiveness, Government of Spain.

KEYWORDS: TV series, social media, TikTok, co-creation, *Bridgerton*, Shonda Rhimes, female empowerment.

REFERENCES

Barlow & Bear (2021). About us. https://www.barlowandbear.com/about

Deci, E. L.; Richard M. R. (1980). The Empirical Exploration of Intrinsic Motivational Processes. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (pp. 39-80). New York, USA: Academic Press.

Erdem, Tulin; Keller, Kevin Lane; Kuksov, Dmitri; & Pieters, Rik (2016). Understanding branding in a digitally empowered world. *International Journal of Research in Marketing*, *33*(1), 3-10. https://doi.org/10.1016/j.ijresmar.2015.12.003

France24. 2020. Covid, pandemia y confinamiento: cómo cambió el mundo en 2020. 15 december. https://bit.ly/2TcxPTL

Greenwald, A. G.; Mahzarin R. B. (1995). Implicit Social Cognition: Attitudes, Self-esteem, and Stereotypes. *Psychological Review*, 102(1), 4-27. http://doi.org/10.1037/0033-295X.102.1.4

Hsieh, S.H.; Chang, A. (2016). The Psychological Mechanism of Brand Co-creation Engagement. *Journal of Interactive Marketing*, 33, 13-26. https://doi.org/10.1016/j.intmar.2015.10.001

Huang, Z. (2021). Leaked ByteDance Memo Shows Blockbuster Revenue Projections. *Bloomberg*. 15 april. https://www.bloomberg.com/news/articles/2021-04-16/bytedance-to-grow-ad-revenue-to-40-billion-ahead-of-ipo

IAB Spain (2021). *Estudio de Redes Sociales 2021*. https://iabspain.es/estudio/estudio-de-redes-sociales-2021/

Liu, C., & Yu, Y. (2020). Inside ByteDance, the $75bn unicorn behind TikTok. *Nikkei Asia*. 20 march. https://asia.nikkei.com/Spotlight/The-Big-Story/Inside-ByteDance-the-75bn-unicorn-behind-TikTok

Marinas, L. (2019). Instagram: Donde Millennials, Generación Z, Mcluhan y Bolter se cruzan. *CIC Cuadernos De Información y Comunicación, 24*, 187-201. https://doi.org/10.5209/ciyc.64641

Padilla-Castillo, G. (2010). *Las series de televisión sobre médicos (1990-2010): tres enfoques: comunicación interpersonal, comunicación institucional, relaciones entre ética, moral y política*. Madrid, Spain: Universidad Complutense de Madrid.

Padilla-Castillo, G. (2021). [TikTok como vía de promulgación de la fe católica]. In Nuria Sánchez-Gey & María Luisa Cárdenas-Rica (Ed.), *Comunicación a la vanguardia. Tendencias, métodos y perspectivas*. Madrid, Spain: Dykinson.

Ramaswamy, V., & Ozcan, K. (2016). Brand value co-creation in a digitalized world: An integrative framework and research implications. *International Journal of Research in Marketing*, 33, 93-106. https://doi.org/10.1016/j.ijresmar.2015.07.001

Tajvidi, M.; Wang, Y.; Hajli, N., & Love, P. E. D. (2017). Brand value Co-creation in social commerce: The role of interactivity, social support, and relationship quality. *Computers in Human Behavior*, 115. https://doi.org/10.1016/j.chb.2017.11.006

Tseng, C.-H.; Kuo, H.-C., & Chen, J.-M. (2013). [The Relationship Among Advertisement, Electronic Word Of Mouth, And Purchase Intention Of Virtual Community Members]. In Avery, A. E. (Ed.), *Proceedings for the Northeast Region Decision Sciences Institute 2013 Annual Meeting* (pp. 129-148). Brooklyn-New York, USA.

Toffler, A. (1979). *The third wave*. New York, USA: Bantam Books.

Visa, M.; Serés, T., & Soto, J. (2018). From the family portrait to the profile picture. Uses of photography in the Facebook social network. *Revista Latina de Comunicación Social*, 73, 718-729. DOI: 10.4185/RLCS-2018-1278.

Verwey, S. (2015). Self-expression and collaborative 'pro-sumption' in the digital brandscape. *Communicatio*, *41*(3), 320-339, https://doi.org/10.1080/02500167.2015.1093327

We are Social y Hootsuite (2020). *Digital 2020 España*. Retrieved from https://wearesocial.com/es/digital-2020-espana

Weimann, G. & Masri, N. (2020). Spreading Hate on TikTok. *Studies in Conflict & Terrorism*, online, 1-14. https://doi.org/10.1080/1057610X.2020.1780027

Zorrilla, M. (2021). Netflix anuncia que 'Los Bridgerton' ha destronado a 'The Witcher' con el mejor estreno de una serie en la historia de la plataforma. *Espinof*, 28 de enero. Retrieved from https://www.espinof.com/series-de-ficcion/netflix-anuncia-que-bridgerton-ha-destronado-a-the-witcher-mejor-estreno-serie-historia-plataforma

@abigailbarlowww (2021). https://www.tiktok.com/@abigailbarlowww?lang=es

TURKISH TV SERIES' SUCCESS IN SPAIN IN A FEMINIST KEY. CASE OF *WOMAN* (*KADIN*)

Aysel Zeynalova, Graciela Padilla-Castillo, Asunción Bernárdez-Rodal

Complutense University of Madrid (Spain)

ayselzey@ucm.es; gracielp@ucm.es; asbernar@ccinf.ucm.es

INTRODUCTION

Mujer (*Kadın*) is a Turkish TV series, broadcast in Spain by Antena 3. It is an adaptation of *Woman*, a Japanese *dorama* written by Yuji Sakamoto. This was one of the reasons why it was thought that the series would not appeal to Turkish audiences, but the audience figures proved the opposite. The episodes of the series broke records for several consecutive weeks, registering between 26.5% and 28.7% share.

The series has three seasons and 81 episodes. The first two seasons have already been broadcast in Spain, with excellent audience figures: always between 10 and 20% audience share, more than 2 million viewers on average in each episode, and leader in its time slot on most of the nights it was broadcast. Its main character, Bahar, is suddenly widowed and struggles to raise her two young children, Nisan and Doruk. Her understanding of motherhood is shaped by her own mother, who abandoned her as a child, and that is why she wants to be a loving, close mother who engages in dialogue.

The aim of the work is to explore the plotlines of the series and its exploitation in Turkey and Spain, taking into accounts the feminist social context of each country, the competition of television series in each grid and their presence in social networks. To achieve this objective, a mixed quantitative and qualitative analysis is used, combining the synopsis with the promotion of the series on social networks. Among the main conclusions, we found that: *Mujer* (*Kadın*) is one of the most feminist TV series in Turkey and triumphs in Spain for telling a universal story that arouses empathy, moving away from the subject matter of other Turkish series, with its marked invitation to women's empowerment on all screens.

ON THE PARADIGM SHIFT IN TELEVISION CONSUMPTION

In June 2020, for the first time, digital advertising investment in Spain exceeded investment in traditional television (InfoAdex, 2020). Of the funds invested in controlled or conventional media, 38.6% went to digital communication. Within the sector, *Display* and *Video* formats (including social networks) received the highest expenditure (InfoAdex, 2020). To these data, revealing a new communication paradigm, we had to add unconventional or below the line communication, which continues to grow exponentially every year and which includes investment in branded content, influencers and native advertising.

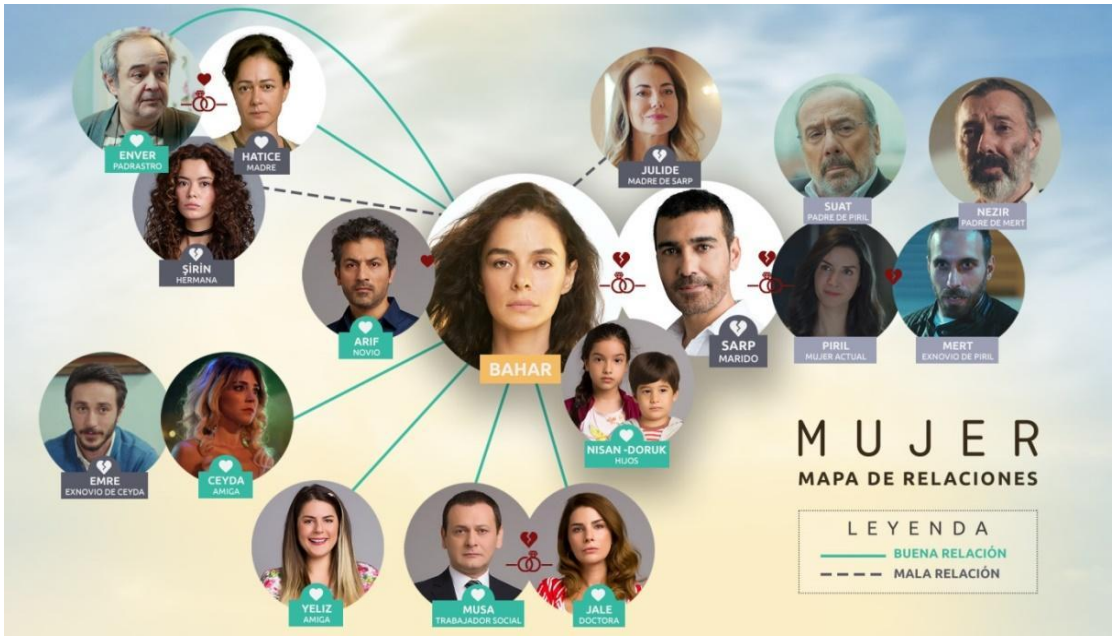
This new environment has been definitively consolidated in March 2021, when InfoAdex has published the complete data for 2020 and has been able to compare January 2020 with January 2021. The *Digital* sector is consolidated in the first position of advertising investment for the entire Spanish territory and is the only one that has not decreased in investment, in the sad year of socio-health crisis and confinement. Within *Digital*, we find *Social Networks*, whose investment figure increased by 2.5%,

reaching 38.7 million euros (InfoAdex, 2021). *Television* is once again the second sector in terms of advertising investment, with 116.4 million euros, which is 19.8% less than the 145.2 million euros in the same period in 2020 (InfoAdex, 2021). *Radio* is in third place, with an investment of 23.3 million euros and a decrease of -26.7%; and *Newspapers*, with 15.6 million euros of investment, recorded a fall of -33.2% on the previous year's figure. Also declining are *Exterior*, with a drop of 62.4%; and *Magazines*, with an investment 33.5% lower than in 2020 (InfoAdex, 2021).

At the same time, Movistar, HBO and Netflix are stealing viewers from traditional television every day with unbeatable prices, extremely and varied catalogues, quality in-house productions and secret algorithms that prescribe the products we might like and detect, in our binge watching, the series that need to be produced to succeed. In order to maintain its space and not fall to the VOD platforms, free-to-air Spanish television is reinventing itself, looking for new products. In this search, Turkish series have stood out in the last three years on Atresmedia and Mediaset channels.

Atresmedia is the communication conglomerate that includes Antena 3, the Spanish TV channel that broadcasts *Mujer* (*Kadın*). Its website (https://www.antena3.com/series/mujer/) offers a multitude of content for the audience: the complete *Chapters* (on the Atresmedia Player platform, with a free subscription), a description of the *Characters*, a selection of *News* (in video and text format) and a compilation of *Best Moments*. In the *Characters* section, the following outline presents the cast and the main relationships of the plot:

Figure 1. Diagram of the characters in *Mujer* (*Kadın*).



Source: https://www.antena3.com/series/mujer/personajes/

As mentioned in the introduction, this proposal studies the progress of the story with its dissemination on social networks, and an important part of this content is also published on the website. Specifically, for this study, we are interested in pointing out the most feminist publications and the numerous news items about this and other Turkish series that Atresmedia has broadcast: *Fatmagül*, *Amor de contrabando*, *Ezel*, *Sila*.... They also announce the acquisition of the rights to more series and arouse the interest of the audience: *Hercai*, *Amor prohibido*, *Las mil y una noches*, *Fugitiva*, *Cennet*, *Mar de*

amores... The proposals are very numerous; they are broadcast on various channels of the platform; and the added content (especially mini-videos) help to maintain the intrigue.

ON THE REINVENTION AND RESURRECTION OF THE SOAP OPERA

The soap opera was adopted in Latin America and gave rise to a new genre. Bermúdez (1980, p. 7) calls it the 'little sister' of the radio soap opera and points out that initially, it was of interest to illiterate people. Forero (2002, p. 103) prefers to bring it closer to the *folletín* because of the common elements of its stories: manichean good and bad, misunderstandings and misunderstandings, rich and poor and a great hidden secret. Martín-Barbero prefers to call it *saga* and explains its beginnings as follows: "In the mid-nineteenth century, popular demand and the development of printing technologies made stories the launching pad for mass production. The osmotic movement was born in the press, a press that in 1830 had begun to move from political journalism to commercial enterprise. This is the birthplace of the pamphlet, the first type of text written in the popular mass format. A cultural phenomenon much more than a literary one, the *folletín* is a privileged space for studying the emergence not only of a means of communication aimed at the masses, but also of a new mode of communication between the classes" (Martín-Barbero, 1995, p. 32).

The same author adds that *folletín* does not come from the popular novel, published by episodes in a newspaper, but that the term indicated a place in the newspaper: "The *basement* of the front page, where the varieties, literary criticism, theatre reviews, advertisements and culinary recipes, and not infrequently news that disguised politics as literature, ended up. What was not allowed in the body of the newspaper could nevertheless be found in the *folletín*, and this condition of origin, as well as the mixture of literature and politics, would leave its mark on the format" (Martín-Barbero, 1995, p. 33). In 1836, the newspaper became a commercial enterprise and the owners of *La Presse* and *Le Siecle*, Parisian publications, introduced advertisements for words and stories written by fashionable novelists (Martín-Barbero, 1995, p. 33). The stories gained space and importance and earned for themselves the name of *folletín*. The author says that the aim was to "reorient newspapers towards the general public by lowering costs and taking advantage of the possibilities opened up by the technological revolution brought about by the rotary press" (Martín-Barbero, 1995).

Beyond the name or the birth, it is necessary to move on to the content of the soap opera. Forero (2002, p. 104) remarks that a soap opera must have a reparation of the *otherness*. This term covers: "The situation in which someone thinks they are one thing and in reality they are another: the girl thinks she is an orphan, but she is the daughter of a millionaire; the bad guy thinks the secretary is his ally, but in reality he is the son of someone he eliminated, who has come to seek revenge; the beggar is in reality a millionaire who looks out of his disguise, and so on" (Forero, 2002). Cabrujas (2002, p. 51) adds that the soap opera is "a system of communication of a human being consisting of a person". It is more than drama because the characters create a whole world. Its protagonist is usually a woman, always a victim, totally or partially. She must appear helpless and have a candid face. She is honest, hates no one and will only break the unjust law for a noble cause (Cabrujas, 2002, p. 56). That cause is precisely the otherness of which Forero (2002) spoke and the basis of the 'dramatic noun', extended in space and time.

Other authors speak of 'conflict' instead of *otherness*. Nash and Oakey (1978, p. 3) defined it as the force of drama. They added that stories should present a problem that the characters suddenly face; and they have to overcome it or be overcome by it. The plot should develop the unfolding of those problems. If that otherness or conflict were resolved in the first chapter, it would be a closed story with an introduction, knot and denouement. The soap opera is far from that, as it lasts for many episodes. It resorts to a *rhizomatic* or flaky narrative structure (Gordillo, 1999, p. 32). The story is

prodigiously dilated, it acquires a multitude of plots and the difficulties are multiplied a hundredfold, even if staggered solutions appear. If the TV series is liked, it is stretched as much as possible. If it is not as successful as expected, the protagonists will soon end up where they belong. However, this solution is very anomalous, as soap operas are designed to have short-, medium- and long-term followers. Their scriptwriters dominate the prolonged discourse because they know how to trap the audience and coil them indefinitely around their desire to follow the story (González-Requena, 1999, p. 121).

To achieve this, soap operas have characters that evolve very slowly and have a rather limited arc. The classic main character is a young girl who is in an environment that is not her own. This makes her vulnerable and a victim of multiple injustices (Forero, 2002). She has to confront an evil man or woman who is the complete opposite of her. However, she has a great incentive: love. Every soap opera protagonist traditionally falls in love with a classic, handsome, educated and cultured leading man from a social class and occupation very different from her. Their love story develops slowly and with many complications, because he is slow to fall in love and has to face all the people around him who disapprove of the relationship, including the evil one. Forero (2002) introduces a third character. She is the housekeeper, fairy godmother and confidant of the protagonist. She knows the *otherness*, the great secret that can lead to the *anagnorisis* or unmasking of the villain or villainess. Her value is fundamental because of her unconditional support for the disgruntled main character. Once the story has been resolved, she usually dies so that the ending is not entirely happy.

Puppo (in Escudero & Verón, 1997) adds some important characteristics of the characters described above: “The characters are feminised and masculinised according to a series of traits that are socially accepted: women are victims and men are perpetrators, men pursue power and women pursue happiness and love, women pursue noble feelings and men pursue sex. Thus there are ambitious and self-interested women who by this trait fall within the domain of masculinity and there are men who do everything for their beloved and who fall within the domain of femininity”. *Mujer* (*Kadın*) comes to break these norms in this female empowerment. From the title itself, we know that the protagonist does not need her husband, who is supposedly dead, and that she alone can get by with two small children.

The result of combining these characteristics of roles and narrative structures, more old-fashioned or more renewed, is the *telenovela* (in Spain and Latin American) or soap opera (the English name). Its audience tends to be predominantly female and it was historically broadcast in the morning or afternoon television slot. *Mujer* (*Kadın*), and all the most successful Turkish series in Spain, also break this paradigm and the channels broadcast them in the evening prime time, with 2 or 3 episodes, spread over one or two nights a week. The structure is always episodic, with 30 to 45 minute instalments and normally two or three advertising blocks per episode. The most important thing is that each episode ends with a shock or unexpected twist, forcing viewers to watch the next episode.

Peñamarín (1995) attributes its long-lasting success to the “flexibility to adapt to new topics of interest and to include the preferences of the audience, the languages, the situations that frame their lives, even altering the script on the fly to incorporate collective events into the plot”. Moreover, “it is important for its success that the serial sounds real, that it speaks a language in common with its audience, that it presents credible situations that can intrigue and involve them, that it is not false, outdated, boring or predictable” (Peñamarín, 1995, p. 12). Beyond its content, the author refers to the universal themes it deals with and adds that the telenovela “dramatizes the conflict between the logic and morality of feelings and the logic and morality of social and practical life” (Peñamarín, 1995, p. 13). We are interested in all these ideas because *Mujer* (*Kadın*) has reinvented some of these concepts and has done so in a feminist key.

TURKISH SOAP OPERA AS A SOFT POWER

Nowadays Turkey still remains as the most liberal Islamic country. Since the establishment of the Republic of Turkey in 1923, Turkey has made significant progress in the development of democracy. At the same time since the conservative islamist party AKP (Justice and Development Party) came into power in 2002, Turkish TV has become more conservative. The official narrative of a supposedly modern, prosperous Turkey is being challenged by this conservative, intolerant backlash. Recently in March of 2021 Turkey decided to withdraw from the 2011 Istanbul Convention widely regarded as the gold standard in international efforts to protect women from violence. Women from all over the country rose up over withdrawal from the Convention. As it can be observed women in Turkey are going through a difficult time.

That's why Turkish soap opera is considered a great tool for female empowerment (Bernárdez Rodal, 2018) and not only in Turkey but also in many other countries. Millions of viewers across the Middle East, North Africa and Europe are hooked on Turkish TV dramas such as *Kara Sevda*, *What Is Fatmagul's Fault?*, *The Magnificent Century*, etc. Middle East women say they find inspiration in Turkish TV dramas. According to Hürriyet (2018), Samira, a victim of sexual harassment in Egypt found the courage to take the perpetrators – army officers – to court after watching Fatmagül, a gang-rape victim in another Turkish drama series, fight for justice. There are also findings indicating the impacts of Turkish TV series not only on women, but also on the flow of daily life in Arab societies.

At a TV program conveyed the following words of a young woman sitting with her male friends at a café in the West Bank: "If there were no Turkish TV series, I would not be able to be here with my male friends. My family has accepted it after watching these series" (TrtHaber, 2021). Blumenthal (1997) considers women empowerment through soap operas as a form of feminist praxis. As could be seen in Brown (1990) she admits the existence of examples of strong and positive women characters in it. Some of the characters portray the life of strong, independent, working women living in a modern way (Padilla-Castillo, 2014; Bernárdez-Rodal y Padilla-Castillo, 2018; Bernárdez-Rodal y Menéndez-Menéndez, 2021). Such positive representation of women on soap operas could act as potential role models of its female audiences.

The success of Turkish shows naturally did not go down well with conservative forces in the Middle East. In Iran, authorities said soaps were "destabilizing the institution of the family" (Hürriyet, 2018) and Saudi clerics issued fatwas against people watching the shows. Despite that, the value of soap opera has increased from a million dollars in 2007 to 130 million in 2012 and by 2023, the Turkish government hopes this will pull in $1 billion from exports.

Turkish series rise in popularity also in Spain. Turkish TV series have created a sensation in our country after being exported to the country for the first time in 2018, said Jose Antonio Anton, the vice president of Atresmedia Group, who is responsible for the import of Turkish series in Spain (*DailySabah*, 2021). As Anton stated in an interview with the Spanish newspaper *El Confidencial* Turkish series played a significant role in helping Atresmedia beat its biggest rival in Spain, Mediaset, for the second time in the last decade (*DailySabah*, 2021). Today, there are over 20 Turkish TV series currently broadcasting on different channels of Spanish television in prime-time slots and receiving the highest ratings.

METHODOLOGY

The aim of the work is to explore the plotlines of the series and its exploitation in both Turkey and Spain, taking into accounts the feminist social context of each country, the competition of television series in each grid and their presence in social networks. To achieve this objective, a mixed quantitative

and qualitative analysis is used, combining the synopsis with the promotion of the series on social networks. İn the promotion we are facing two different markets: Turkey, on one hand and the Spanish market, on the other hand. This allows us to carry out a clear analysis of how the series engages with users, a comparison between established and relatively new social media strategies used in both countries to promote content and assessing if different cultural contexts impact this strategy. In order to do so, data from all social media accounts of a series is retrieved and analyzed.

BACKGROUND OF THE SERIES

Kadın is a drama television series based on the 2013 Japanese drama *Woman-My life for my children* created by Yuji Sakamoto for Nippon TV. The series was broadcast on the Turkish channel FOX every Tuesday between 2017 and 2020, and was produced by Med Yapım and MF Yapım. In Spain the series was first broadcast by Antena 3 only 1 day a week and now it is broadcast 3 days a week, from Monday to Wednesday.

The Turkish version of *Kadın* has 81 episodes over 3 seasons with duration of 120 minutes for each episode. It was broadcast to 65 countries, most of them in Middle East and Turkic speaking Asia but also Balkan and South Europe. The show achieved excellent audience figures in Turkey: always between 26% and 30% audience share and became the leader in its time slot on most of the nights it was broadcast. Its final episode was one of the most viewed episodes in Fox TV's history.

Kadın has succeeded in removing from leadership positions other series broadcast the same day in other TV channels like *Ramo* (Show TV), *EDHO* (ATV) and *Hekimoğlu* (Kanal D). The series has won a number of awards: a special award at the Tokyo Drama Awards and the awards for "Best Actress" for Özge Özpirinçci and "Best Child Actor" for both Doruk and Nisan (Kübra Süzgün and Ali Semi Sefil) at the Golden Butterfly Awards. These awards are distributed at a highly anticipated annual ceremony in Turkey that rewards the world of television and music in that country. .

İn Spain the impact went beyond all expectations. *Kadın* has beaten all records by reaching an audience of 2.141.000 and 18% share in the prime time slot. The popularity of Turkish series in Spain is also due to the nature of the plot and the quality of the shooting of Turkish film production. In general terms, there is something exotic, yet at the same time quite familiar, to the audience. Viewers in the Arab countries see the Turkish woman as a model of the modern Muslim female. The most interesting thing is that while women in Middle East says to be inspired by the modern, feminist narrative of Turkish dramas, the Europeans says that Turkish series have triggered a nostalgia for ideals such as tradition and family ties.

PLOT AND CHARACTERS

Bahar is a young widow with two children, Nisan (7) and Doruk (3). She was abandoned by her mother Hatice when she was only 8 and later loses her father and paternal grandparents. A few years later Bahar meets Sarp and marries him. They lived some years together until Sarp's death in a ferry accident. Bahar is a devoted mother and tries hard to provide for her beloved children. She tries desperately to hide from Nissan and Doruk the state of misery and precariousness in which they live creating an imaginary universe: she tell her children that they will live in a palace (Saray: the name of the building where they move which means palace in Turkish) and tell them that their next-door neighbour Ceyda, a cabaret singer with long blond hair is a princess. When Ceyda is beaten, she tells the children that this is the shooting of a new TV series.

Bahar worked as an ironer in a textile workshop. Upon her friend's Yeliz suggestion, Bahar goes to the social welfare office in order to get social assistance from the government. Then an employee at the office asked if she has any relative who can support her, Bahar says no. Her father passed away years ago and she has not been seeing her mother for a long time. After that, an employee tries to contact Hatice. He learns that she is willing to help Bahar. Bahar decides to meet her mother and over time wants to get close to her again but her half-sister Şirin tries to prevent her mother and sister from rekindling their relationship. In order to protect Şirin's mental well-being, Hatice turns her back to Bahar. Then Enver, Hatice's husband, decides to help his step-daughter. And so do Yeliz, Ceyda and Arif, the son of the owner of Bahar's flat.

Meanwhile, Bahar starts to suffer from pain and fainting and learns that a bone marrow transplant is needed in order to save her life. Soon everyone learns about Bahar's condition and finds out that Şirin's bone marrow is suitable for transplant to Bahar. Meanwhile, Sarp is shown to be alive. Now he is a rich man married to Pırıl and has twin children. Pırıl's father, Suat tells Sarp that Bahar and her children died 4 years ago in a terrible car accident, which is a lie made up by Şirin. To prove this, Suat even makes fake graves and shows them to Sarp. One day Sarp meets Şirin in a grocery store. It's revealed that Şirin was in love with Sarp 4 years ago and she was the main responsible for Sarp's ferry accident.

Bahar's condition worsens and she accepts to go to hospital. Arif has feelings for Bahar, and he proposes to her. One day, Doruk sees Sarp across the road and calls him father. Sarp is shocked and decides to go and see Doruk at Enver's house. Enver tells Sarp that Bahar and his kids are alive. Then Sarp learns about Bahar's situation and brings Şirin to give bone marrow to Bahar. After Bahar's recovery, she meets Sarp after being separated for years. She accepts Sarp, unaware that he has remarried. Once she learns this, Bahar tries to ignore him. One night, Sarp comes to pick Bahar and his children up to go to Pırıl's house, but they are followed by murderers. Sarp escapes with Bahar and their kids from Bahar's home when the murderers surround it. Unfortunately, Yeliz is shot and dies.

One day, Bahar faints on the road and Hatice, Sarp and Arif try to bring her to the hospital. While Arif was driving the car, he got into an accident. The four of them get injured. Bahar recovers, but Hatice dies. Şirin comes to see Bahar and Sarp to the hospital and closes the intravenous flow rate of the line connected to Sarp and kills him. Then she blames Arif for the death of her mother and Sarp. Arif gets imprisoned.

3 months later, Bahar is living peacefully with her children in their house. Arif is released from jail. Enver and Şirin move into a new house in Bahar's apartment after their home burns down. Şirin tells Nisan and Doruk that Arif had killed Sarp and Hatice on purpose. Nisan and Doruk ignore Arif but after learning the truth they change their mind.

Meanwhile Bahar's friend Ceyda gets a career job at the house of Fazilet, a famous writer. She takes care of Raif, Fazilet's son. Fazilet meets Bahar and invites her to share her story. Fazilet writes Bahar's story, naming the book *Kadın* (Woman).

Soon, Arif finds that Sarp was killed by Şirin. Once Bahar learns this, Şirin gets arrested by the Police, who admit her to a mental hospital. Meanwhile, Arif and Bahar's relationship gets stronger when Raif and Ceyda fall in love. Bahar and Ceyda plan to have their marriages on the same day. Bahar, Arif, Ceyda, Raif, Enver, Nisan, and Doruk celebrate together happily.

WOMEN REPRESENTATION IN *KADIN*

After analysing the role of women in both Spanish and Turkish TV we can see that there is a persistence of males playing the leading roles, as well as a reiteration of female stereotypes. Different studies

demonstrate the perpetuation of gender clichés amongst which the following stand out: the representation of men as dominant and women as complementary, less ambitious and dedicated to caring for others, and the tendency to relegate female characters to the private field, leaving the field of work as a basically for males (Padilla-Castillo, 2009; Padilla-Castillo, 2014; Bernárdez-Rodal, 2015; Bernárdez-Rodal, 2018; Bernárdez-Rodal & Padilla-Castillo, 2018; Padilla-Castillo & Sosa-Sánchez, 2018).

When women are shown in the role of the professional woman and the single mother, the focus on these latter seems to be based on patriarchal ideological vision: being wife and mother is seen positively, but woman power is the villain. This situation limits women's identity and imposes certain behavioral patterns on women. And also there is a dichotomous representation in the relationship between woman and motherhood. While motherhood is a result of a woman's biological life; suddenly it becomes a supra-identity that dominates and manages femininity (Padilla-Castillo, 2009; Padilla-Castillo, 2010; Bernárdez-Rodal & Padilla-Castillo, 2018; Ortega-Fernández & Padilla-Castillo, 2020). Meanwhile a man can be a good father, a good husband, and a successful one. The woman should adopt certain roles: housewife, mother, wife and only these roles are compatible with the expectations of the society.

One of the reasons *Kadın* draws attention is because of a significant presence of female characters playing the leading role in the series. Most of Turkish dramas are mainly written by female scriptwriters who nudge the narratives into more feminist paths, and sometimes even attempt to involve their audience. For instance in the case of Fatmagül, when the final court scene of *What Is Fatmagul's Fault?* was filmed, the extra cast to carry banners and shout slogans in support of Fatmagül were real-life victims of sexual abuse.

İn *Kadın* the main character Bahar is shown not only as a wise mother, good wife, devoted to the well-being of the family but also as a very strong, autonomous woman that refuses to marry another man in order to get a better life. But In the series, although the struggle of women to exist in society on their own feet seems to be the basic material of the narrative, it should be emphasized that the dominant ideological discourses imposed by gender codes are implicitly reproduced. Because the point to be considered here is the fact that the narrative of the series implicitly states that "a woman cannot exist without a man and that she will never be complete even if she tries to stand up". Turkey needs more female character role models who don't base their identity on men.

In the series we also can see doctor Jale's story. In the season 3 she ended her unhappy marriage by her own decision and started to live with her son. She tries to be a good mother while at the same time developing a meaningful and successful career, helping her patients and especially Bahar to find a suitable marrow. In the last season we can observe that she is almost the only female character who continues her life on her own way without the need for a male figure to be "complete" and feel "happy". She is definitely the most feminist character in the series and so often speaks out feministic ideas and questions many traditions imposed by the Turkish patriarchal society.

Kadın also offers us a great opportunity to see a true, supportive women friendship. The relationship between Ceyda, a prostitute from a city's cabaret and Bahar is very special. It is quite striking that in a Turkish series a prostitute has a relevant role with a "happy ending" and above all, she is being shown to us from her most human side. As we can see Ceyda suffers abuse and mistreatment of her boyfriend who also scares Bahar and who comes to refer to Ceyda as his "property". At one point in the series, Ceyda decides to break up with him and he asks Bahar for help to get back to her. The response of Bahar is the most feminist answer that she could gave in a society where such type of violence is consented and silenced in most cases: "I don't want my children to grow up watching a man hit women for any reason I don't want men to hit women for any reason".

PROMOTION OF A SERIES IN TURKEY AND SPAIN

The series that began to be broadcast in prime time in Turkey on October 27, 2017 not only had the standard promotion through social networks and previews of the episodes on TV, website, YouTube channel but also was promoted through the participation of the entire cast in different solidarity projects like blood donation for LÖSEV (*Lösemili Çocuklar Vakfı* or Foundation for children with leukemia), collaborations with different brands such as Finish through advertising campaign on TV. The actress Özge Özpirinçci has also participated in projects against domestic violence. Turkish actresses are fully conscious of what it is that they are doing and the impact it has. Many of them have taken the effort outside the TV studio by participating in a campaign to stop domestic violence against women. It's not just a marketing strategy. Some of them genuinely believe they can help.

Kadin started broadcast on July 7, 2020 reaching 12.7% screen share, and on February 9 reached a maximum of 23.7%. During its first episodes and due to summer programming, Antena 3 decided to broadcast full episodes but in September 2020, with the start of the television season, the channel decided to divide the chapters into two parts so that their end would not be later than 00:45. It was the second most watched series in 2020 on Spanish television, behind *Mi Hija* (*Kızım*). The series is also available on AtresPlayer, subscription based streaming platform of Atresmedia.

PROMOTION IN SOCIAL MEDIA

Nowadays social networks have become one major element for media production companies to communicate with their prospects and clients (Fernández-Gómez & Martín-Quevedo, 2018). The series *Kadın* is present on the following social networks: Twitter, Facebook and Instagram, but the main difference is that in Turkey the series has its own account on each social network, while in Spain the content is shared from the Antena 3 account on these social media.

A content analysis is carried out to distinguish the type of content posted on different social media accounts (if it is original content or user-generated as well as the type of content itself, this is, promotional, informative or external and the intention of this content which can be to inform, humorous, or call-to-action.

In general, Turkish accounts on Twitter, Instagram and Facebook have shown a tendency of being dominantly humorous, sharing funny moments from the series and from backstage, as well as trying to engage and interact more with users through calls-to-action. There is a creative approach to promote and inform about series through social media language, such as hashtags created for each episode in particular and visual resources like videos, Throwback Thursday backstage photos, mini-videos from main characters inviting people to watch the show and comment on it in social media.

It should be noted that most of these hashtags created especially for each episode of the series became a trending topic on Twitter during the first season broadcast (Televizyon Gazetesi,2017). As for the mini videos or photos published, the ones that appear the most are Bahar, Sarp, Arif, Ceyda, Raif and Şirin. The friendship and sisterhood between Bahar and Ceyda is one of the most frequently mentioned topics in the account. Besides that, there are posts about announcements of different events that were held during the broadcast of the series such as aforementioned blood donation for children with leukemia, donation to autism research or charity events to support victims of domestic violence.

As we have mentioned before, in the case of Spain, the series does not have its own accounts on the social networks and everything is shared through the Antena 3 accounts. The character of the posts is promotional and informative. Likewise, mini videos of 1-2 minutes are shared where the web-editor

Carlota Galdon talks about the characters of the series, the past chapters and other curiosities related to the series. Many Turkish actors including the protagonists of the series know that Turkish series are seen a lot in Spain and they send dedicated messages to their Spanish fans. These messages are shared in Antena 3 social media accounts. As we see in the case of Spain it is more focused on posting interviews, photos of actors from their personal social media accounts, information about how the series has been shot, etc. All these content especially videos help the channel to maintain the intrigue and keep the audience interested in knowing more and more.

The fact that both markets use their platforms to predominantly share promotional content indicates great efforts towards a strong brand marketing strategy, presenting a consistent identity through different social network platforms.

However, there is also significant difference between both markets particularly in terms of intention when posting the content. While Spanish accounts tend to be more informative about the content that they post (usually information about episodes), Turkish accounts share more content with the intention of being humorous. This difference happens on Twitter, Facebook and Instagram, although the latter platform is even more highlighted.

ACKNOWLEDGEMENTS

This research has been carried out as part of the competitive research project: “Produsage cultural en las redes sociales: industria, consumo popular y alfabetización audiovisual de la juventud española”. R&D project of the Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia 2018-2022. Reference: FEM2017-83302-C3-3-P. Ministry of Economy and Competitiveness, Government of Spain.

REFERENCES

Bernárdez, Rodal, A. (2015). *Mujeres en medio(s): propuestas para analizar la comunicación masiva con perspectiva de género*. Madrid: Fundamentos.

Bernárdez-Rodal, A. (2018). *Softpower. Heroínas y muñecas en la cultura mediática.* Madrid: Fundamentos.

Bernárdez-Rodal, A. & Menéndez-Menéndez, I. (2021). Ageing and the Creative Spirit of Women in the Audiovisual Market: The Case of Olive Kitteridge (2014). *International Journal of Communication*, 15, 563-580. https://doi.org/1932–8036/20210005.

Bernárdez-Rodal, A. & Padilla-Castillo, G. (2021). Mujeres cineastas y mujeres representadas en el cine comercial español (2001-2016). *Revista Latina De Comunicación Social*, 73, 1247-1266. https://doi.org/10.4185/RLCS-2018-1305.

Bermúdez, M. (1980). *La ficción narrativa en radio y televisión*. Caracas: Monte Ávila

Blumenthal, D. (1997) *Women and Soap Opera: A Cultural Feminist Perspective*. Connecticut: Greenwood.

Cabrujas, J. I. (2002). *Y Latinoamérica inventó la telenovela*. Caracas: Alfadil.

DailySabah (2020). *For Spaniards and Russians, Turkish TV series go beyond passion*. Retrieved from https://www.dailysabah.com/arts/cinema/for-spaniards-and-russians-turkish-tv-series-go-beyond-passion

Escudero, L. & Verón, E. (Eds.) (1997). *Telenovela. Ficción popular y mutaciones culturales*. Barcelona: Gedisa.

El Confidencial (2020) *¿Cómo hacerse con un bombazo? Así cerró Antena 3 la compra de 'Mujer'* Retrieved from https://www.elconfidencial.com/television/series-tv/2020-12-14/proceso-compra-mujer-serie-turca-antena3_2855675/

Forero, M. T. (2002). *Escribir televisión. Manual para guionistas*. Barcelona: Paidós.

González-Requena, J. (1999). *El discurso televisivo: espectáculo de la posmodernidad*. Madrid: Cátedra.

Gordillo, I. (1999). *Narrativa y televisión*. Sevilla: Mad.

Hürriyet (2019) *Türk dizileri için Türkçe öğreniyorlar! Araplar yasak dinlemedi*. Retrieved from https://www.hurriyet.com.tr/kelebek/magazin/araplar-yasak-dinlemedi-41087871

InfoAdex (2020). *Estudio InfoAdex de la inversión publicitaria en España 2020.* Retrieved from https://www.infoadex.es/home/wp-content/uploads/2020/02/Estudio-InfoAdex-2020-Resumen.pdf

InfoAdex (2021). *Estudio InfoAdex de la inversión publicitaria en España 2021.* Retrieved from https://www.infoadex.es/home/wp-content/uploads/2021/02/PDF-2021-Presentaci%C3%B3n-Estudio-MADRID.pdf

Martín-Barbero, J. & Rey, G. (1999). *Los ejercicios del ver. Hegemonía audiovisual y ficción televisiva*. Barcelona: Gedisa.

Nash, C. & Oakey, V. (1978). *The television writer's handbook. What to write, how to write it, where to sell i*t. New York: Barnes & Noble Books.

Ortega-Fernández, E. A. & Padilla-Castillo, G. (2020). "Diálogo transmedia de las series de televisión. La superación de la Quinta Pared en *House of Cards*". *Revista Estudios sobre el Mensaje Periodístico*, 26(3), 1101-1120.

Padilla-Castillo, G. (2009). La mujer en el cine de Kenji Mizoguchi. *CIC Cuadernos De Información Y Comunicación*, *14*, 251-267. Retrieved from https://revistas.ucm.es/index.php/CIYC/article/view/CIYC0909110251A

Padilla-Castillo, G. (2010). Las series de televisión sobre médicos (1990-2010). Su éxito desde el Análisis Transaccional y la Ética (I). *Revista de análisis transaccional y psicología humanista*, *63*, 244-260. Retrieved from http://com.aespat.es/Revista/Revista_ATyPH_63.pdf

Padilla-Castillo, G. (2014). Teoría de la información y de la comunicación en una serie de televisión: scandal. *Historia Y Comunicación Social*, *19*, 133-144. https://doi.org/10.5209/rev_HICS.2014.v19.45016

Padilla-Castillo, G. & Sosa-Sánchez, R. P. (2018). Ruptura de los estereotipos de género en la ficción televisiva sobre el poder político: el caso Borgen. *Vivat Academia*, 145, 73-95. https://doi.org/10.15178/va.2018.145.73-95

Peñamarín, C. & López-Díez, P. (Eds.) (1995). *Los melodramas televisivos y la cultura sentimental*. Madrid: Dirección General de la Mujer e Instituto de Investigaciones Feministas de la Universidad Complutense de Madrid.

3. Ethical Trends and Technological Opportunities after Covid-19

"YOU MUST HAVE YOUR WEBCAM ON FOR THE ENTIRE DURATION OF THE EXAMINATION": THE TRADE-OFF BETWEEN THE INTEGRITY OF ON-LINE ASSESSMENTS AND THE PRIVACY RIGHTS OF STUDENTS

Damian Gordon, J. Paul Gibson, Brendan Tierney, Dympna O'Sullivan, Ioannis Stavrakakis

Technological University of Dublin (Ireland), Institut Mines-Télécom (France), Technological University of Dublin (Ireland), Technological University of Dublin (Ireland), Technological University of Dublin (Ireland)

Damian.X.Gordon@TUDublin.ie; Paul.Gibson@Telecom-SudParis.eu; Brendan.Tierney@TUDublin.ie; Dympna.OSullivan@TUDublin.ie; Ioannis.Stavrakakis@TUDublin.ie

ABSTRACT

The impact of COVID-19 has been widespread and far-reaching, and one domain that has experienced severe disruption is the university education sector, where the entire apparatus of teaching and assessment for many programmes of study had to move on-line in a matter of days[10]. This was accomplished notably through enormous co-operation between staff and students in educational institutions (Adnan and Anwar, 2020). The negative economic impacts of COVID-19 on university students has been highlighted in terms of poor access to online resources, delayed graduation and lost internships with this effect felt more keenly by students from low socioeconomic backgrounds (Aucejo, et al. 2020). However, an issue that has been less reported is how the crisis highlighted mismatches between on the one hand the regulations and requirements of the educational institutions, and on the other hand the privacy rights (and needs) of the students.

In this research we are investigating the challenges associated with the potential for students and teachers to inadvertently share aspects of their private lives as part of on-line teaching and assessment, as well as the ethical challenges of monitoring students during exams. Some educational institutes have used software for monitoring students during assessments (called *e-Proctoring systems* (González-González, et al., 2020)), and these systems lead to a range of potential ethical concerns, particularly if the systems employ facial detection (or recognition) systems and/or artificial intelligence systems to detect potential malfeasance.

One voice that hasn't been included in this discussion heretofore is the student voice, so this research includes the design and development of the *WebCam Usage Student Survey* (*WUSS*), and a group of computer science students (N=44) were asked for their opinions on a wide range of privacy issues (as these students have some idea on the potential pitfalls of using these types of technologies). Their views are varied and nuanced, and their perspective in combination with the literature provide a complex picture of the ethics of online interactions.

This issue is one of a rapidly growing number of computer ethics issues that have been emerging recently, to such an extent that a number of third-level institutes across Europe are collaborating to explore some of these key ethical challenges, and to develop educational content that is both based

[10] https://www.timeshighereducation.com/hub/keystone-academic-solutions/p/impact-coronavirus-higher-education

on pedagogically sound principles, and motivated by international exemplars of best practice to highlight these matters as part of the Erasmus+ Ethics4EU project[11] (O'Sullivan and Gordon, 2020).

INTRODUCTION

Given the abrupt nature of the move to on-line teaching that was dictated by COVID-19, educational institutions were not necessarily in a position to fully consider the ethical ramifications of their decisions, or to update their policy documents. Many were also unable to obtain so-called "wet signatures" for explicit consent forms from students for this new approach, or for the use of e-Proctoring systems (González-González, et al., 2020). Student groups and Digital Rights advocates have begun to raise significant concerns about these systems, and the mandatory use of webcams in on-line teaching and assessment. A news article by Nir Kshetri in "The Conversation" on November 6th, 2020[12] points out that in America organisations such as the Electronic Frontier Foundation have filed numerous petitions to academic institutes and legislative bodies to call for educational administrators and teachers to end the use of these systems, and categorised their use as "spying".

Some have argued that the best way to deal with these issues is to avoid them altogether, so to have neither students nor teachers to ever turn on their webcams during lessons, and to change the type of assessment to one that doesn't require invigilation, for example, using open-book examinations which are mainly focused on applying knowledge as opposed to assessing basic recall (Remtulla, 2020). In situations where this is possible, it is a viable approach, although since the introduction of the Bologna process in 1999 (which impacted higher education in 29 European countries), with its emphasis on learning outcomes, it is more challenging to develop more open and individualised assessment approaches (Murtonen, et al., 2017; Zeide and Nissenbaum, 2018).

If students and teachers are required to share their webcams, this may inadvertently lead to them sharing aspects of their private lives as part of on-line teaching and assessment, this could include sharing visual information about their private residences; or sharing audio information that might reveal too much information about their private lives. On the other hand, some teachers feel it is difficult to foster a connection with their students without seeing their faces, and encourage students to share their webcams, this can sometimes unintentionally cause students to feel anxious (a particular concern for students appears to be concern over their peers' perceptions of them (Rajab and Soheib, 2021). Further issues might arise if the staff or students are *required* by their educational institute to always have their webcams on during lessons or assessments. This can blur the differentiation between public spaces and private spaces, which philosophers like Jürgen Habermas (1991) and Hannah Arendt (1998) have explored through questions of ownership and property, and they asked questions such as; "Who owns resources in these spaces?" and "What is truly private?" There are also a number of other "divides" worth exploring: race, social status, gender, etc. For example, in the context of gender, female students and staff tend to be more cautious about sharing their webcams, as they are more likely to be harassed and exposed to aggressive behaviours in an on-line setting (Chawki and el Shazly, 2013).

Educational institutions that require students to use webcams to be active during online assessments often use software called e-Proctoring systems to monitor the activities of the students during the assessment process. These systems replace a human invigilator (or *proctor*) who ensures that all of the necessary examination regulations are adhered to, and help to prevent cheating in a brick-and-mortar

[11] http://ethics4eu.eu/

[12] https://theconversation.com/remote-education-is-rife-with-threats-to-student-privacy-148955

educational setting. There are a growing number of such systems available, such as Remote Proctor NOW (RPNOW), eProctoring, SMOWL and ProctorExams (González-González, et al., 2020), and these e-Proctoring systems typically can be either manual or automated, where manual proctoring (also known as *Live Proctoring*) is remote invigilation where a person is actively supervising the test-taker throughout the assessment, whereas automated proctoring uses technologies such as machine learning and facial detection to monitor both the test-taker and their technologies, including laptops, tablets, and mobile phones. These systems raise a number of security and fairness considerations (Langenfeld, 2020), additionally at least one of these systems have trouble detecting persons-of-colour[13]. It is worth noting that students do not always have full control over the environment in which they take their examinations, whether in student residences or in a family home, if someone enters the room that they are in, or a noise is heard in the background, some of the automated systems will log the student out, and others will even summarily fail them. Some e-Proctoring systems enforced these automated processes and others do not, so it is important that students and teachers be fully aware of the conditions and consequences of using these systems rather than allowing potential misinformation about the functions of these systems to increase their anxiety. In fact, De Santis, et al. (2020) found that students who have used e-Proctoring systems previously (whether automated or manual) are significantly more confident with their use for assessment purposes.

Some of the issues around student anxiety appear to originate from concerns around surveillance, and from a philosophical perspective such systems cannot fail but bring to mind the notion of a *Panopticon*, a building design (and a system of control) that allows all people in that building to be observed by a single, central observer. Developed by English philosopher and social theorist Jeremy Bentham in the 18th century, the concept has been viewed as the blueprint for a tool of oppression and social control by philosophers like Michel Foucault (1977) and Gilles Deleuze (1992), who see such systems as a means of control by groups of people (including students) through disciplinary power. Allen (2012), Tufekci (2017), Zuboff (2019) and Vatcha (2020) further explore the nature of digital surveillance, and such considerations should be incorporated into the decision-making processes of educational institutes when they are considering the use of e-Proctoring systems.

Another area of concern is that a minority of these systems require that students display some form of identification (e.g. passport or driving license) to validate the initial system login process, this represents a significant security concern, as it is possible in some of these systems for third-parties to intercept the video and audio information being transmitted (notably, intruders have been able to gain access to Zoom classrooms - known as "Zoombombing" - due to issues with Zoom's cybersecurity). This leads to a range of serious questions about the recording and retention of this data, and particularly around the issue of ownership of that data. Even if it were possible to establish legally by whom the data is owned (potentially the students, the platform suppliers, the educational institution, or some combination of these stakeholders), the ethical ownership of this data is far less clear. A concomitant consideration is around the issue of consent; how can it be given if the ownership of the data is difficult to establish, and how can it be meaningful if it isn't clear how this data will be used in the future? In general, the use of automated machine learning and facial detection techniques in any computer system should be viewed as a matter of concern, especially since on 30th June 2020, the Association for Computing Machinery (the professional body for computer professionals) called for the cessation of all use of facial recognition technologies, as they produce "*results demonstrating clear bias based on ethnic, racial, gender, and other human characteristics recognizable by computer systems*" (ACM, 2020). Andrejevic and Selwyn (2020) examined the issue of facial recognition in the

[13] https://www.theverge.com/2021/4/8/22374386/proctorio-racial-bias-issues-opencv-facial-detection-schools-tests-remote-learning

educational context, and raised concerns around the dehumanising nature of this technology, which can lead to the foregrounding of gender and race, as well as concerns around the dangers of using the data from these systems in automated decision support systems.

Researchers S.E. Eaton and K.L. Turner (2020) highlight concerns about the relationship between e-Proctoring systems and student mental health, and conclude that more research needs to be done to explore their relationship. Regehr and McCahan (2020) note that e-Proctoring systems have been used to an unprecedented level during the COVID-19 crisis, which has resulted in a number of scalability and technical challenges, including connectivity issues for students, which has contributed significantly to their stress levels, and opens up the possibility of sharing exam questions between students taking the same examination at different times. Coghlan, et al. (2020) philosophically analyse e-Proctoring systems, and they highlight some of the dangers of these systems, such as in one case when a student's credit card details were accidentally displayed on their computer screen. They conclude that educational institutes must be accountable when mistakes occur, but that the students also bear some responsibility for their choices.

METHODS

An important element that seems to be missing from much of the research heretofore is the inclusion of students' voices in the analysis, therefore this research was designed to incorporate their contributions to this debate. To achieve this, a new survey instrument, the *WebCam Usage Student Survey* (*WUSS*), was developed, inspired by a number of questionnaires related to on-line privacy, and particularly the *Privacy Attitudes Questionnaire* (PAQ) by Chignell, et al. (2003), as that instrument most closely aligns to the goals of this research. It has a number of Likert scales (from *Strongly Disagree* to *Strongly Agree*) that relate to different categories of privacy, and for this research we took the nine questions from the PAQ that relate to the category of *Willingness to be Monitored*, as a springboard for the development of our instrument. Those questions were:

1. I frequently would like to block my phone number on call display
2. I respond to telephone marketing surveys
3. I prefer not to have my name listed on a building directory
4. I would give my home phone number to business clients
5. I like to fill out surveys and contests
6. Red light (intersection) cameras should be used
7. Speeding cameras should be used
8. Insurance companies should not have access to people's health records
9. CCTV should be used in public places to improve public safety and security

However, Question 4 was changed from "business clients" to "lecturers" to make it more applicable to students. Following this, a number of questions were developed and trialled to look more specifically at issues related to webcam usage, and ultimately seven more questions were added to the questionnaire using the same structure and phraseology as the PAQ, as follows:

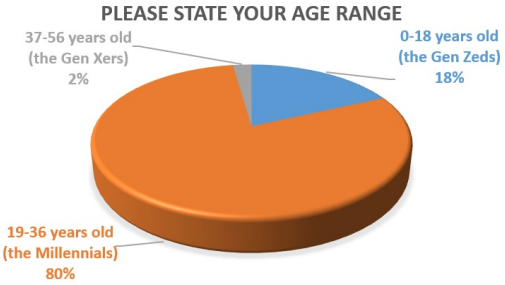
1. I use privacy software or incognito browsing to protect my privacy online
2. I have used the (sliding) camera cover to block the webcam, or have blocked the camera in some other way.
3. It should be mandatory for students to have their webcams on during class
4. It should be mandatory for students to have their webcams on during exams
5. Facial recognition software should be used with the students' webcams to ensure the right person is doing the exam
6. Artificial Intelligence systems should be used with the students' webcams to log the student out of the exam if the system thinks they are doing something suspicious
7. I treat the webcams on my laptop, tablet, and mobile phone in the same way, in terms of privacy considerations

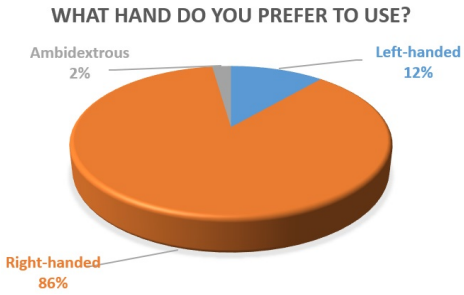
Additionally, demographics questions were added to explore if there are any disparities in the perspectives of different groups of students, based on surveys by Kezer, et al. (2016) and Umawang (2019). The additional questions enquired the students age ranges, handedness, gender, county of origin, primary language, and whether they wear corrective lenses. These are as follows:

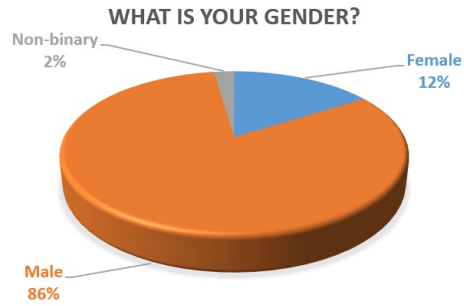
1. Please choose your age range
2. Do you wear corrective lenses (glasses, contact lenses, etc.)?
3. What hand do you prefer to use?
4. Is English your first language?
5. Country of origin (optional)
6. What is your gender?

The survey was given to a range of students enrolled in computer science programmes (both undergraduate and postgraduate). Because the students already have some understanding of both the benefits and pitfalls of the technologies associated with this scenario (for example, Artificial Intelligence, Machine Learning, Image Processing, and Computer Vision), it was felt that they would be able to offer an informed opinion on these matters. It was created using Microsoft Forms, and was distributed from April 21st to April 26th, 2021. The students were given the following key instructions: that the survey is voluntary, that all submissions do not record the students' names, and that the results will be published as part of the broader discussion on these issues. A total of 44 students participated in the survey, and Table 1 presents the demographic results of those participants.

Table 1. Responses to Demographic Questions.







Countries of Origin:

Ireland: 22 participants

Philippines: 4 participants

China: 2 participants

India: 2 participants

Poland: 2 participants

England: 1 participant

Malaysia: 1 participant

Malta: 1 participant

Romania: 1 participant

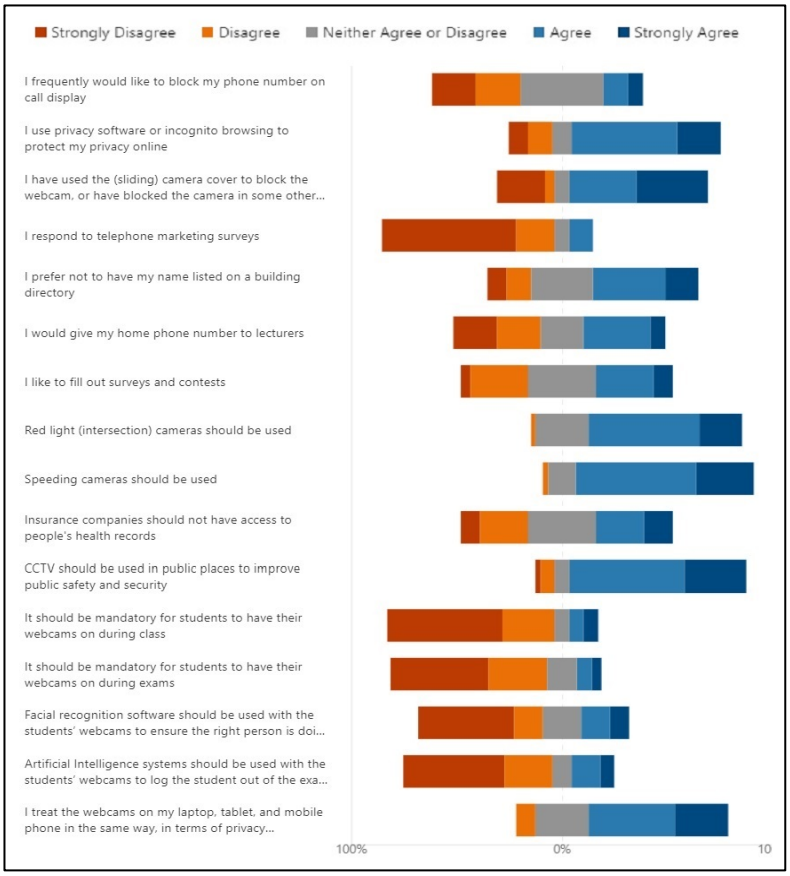
Syria: 1 participant

Vietnam: 1 participant

As would be expected from Computer Science studentgroups, the majority of respondents are male, and principally millennials (in the age range 19-36 years old). There is a reasonable distribution of those who wear corrective lenses, and those who don't, as well as those for whom English is their first language, and those it isn't. The participants represent students from 11 countries, with the majority from Ireland (the country where the survey was conducted).

Following these questions, the rest of the survey was concerned with presenting the privacy scenarios developed from the combination of the PAQ questionnaire and the questions added for this research. Table 2 presents the results of those questions.

Table 2. Responses to Privacy Scenarios.



There are several noteworthy outcomes from this portion of the survey, but the most important overriding message is that there is no scenario that the students are completely unanimous about; although there are some scenarios that the majority of students show some agreement on.

Scenarios where the majority of students either *Agreed* or *Strongly Agreed*, include "*I use privacy software or incognito browsing to protect my privacy online*" at a rate of 70.5%, "*I have used the (sliding) camera cover to block the webcam, or have blocked the camera in some other way*" at 65.9%, and "*I treat the webcams on my laptop, tablet, and mobile phone in the same way, in terms of privacy considerations*" also at 65.9%. These three results combined would tend to suggest that students are generally concerned about their privacy in their private spaces. And these are further supported by the following answers that students also either *Disagreed* or *Strongly Disagreed* with "*It should be mandatory for students to have their webcams on during class*" (79.5%), "*It should be mandatory for students to have their webcams on during exams*" (74.4%), "*Facial recognition software should be used with the students' webcams to ensure the right person is doing the exam*" (59.1), and "*Artificial Intelligence systems should be used with the students' webcams to log the student out of the exam if the system thinks they are doing something suspicious*" (70.4%).

In contrast to their views on private spaces, the students were far less concerned about their privacy in public spaces, for example, "*Red light (intersection) cameras should be used*" (students either *Agreed* or *Strongly Agreed* at a rate of 72.7%), "*Speeding cameras should be used*" (84.1%), and "*CCTV should be used in public places to improve public safety and security*" (84%). These three results combined would tend to suggest that students are generally less concerned about their privacy in public spaces.

It is worth noting that there was no significant difference found in responses between different demographic groups amongst the participants, but this may be due to the sample size, as was noted already, the majority of respondents were male and in the millennial age range. It is also worth noting that research has consistently shown that millennials are confused on the topic of privacy, for example, a study by the USC Annenberg Center for the Digital Future and Bovitz Inc.[14] showed that although a majority of respondents agreed no one should have access to their data or online behaviour, 25% of them said they would exchange information for relevant advertising, 56% would share their location for coupons or deals, and 51% said they would share information with companies if they get something in return.

DISCUSSION

The purpose of this study is to explore ethical issues around the use of webcams and e-Proctoring systems, but not to portray these systems as being inherently problematic, nor is it intended to criticise the developers of these systems. At a time of global pandemic, it became necessary to change how teaching and assessment occurred, and educational institutions have been doing their best to fulfil their obligations to their students. Educators have been finding new ways to teach in these changed circumstances, and ways of connecting with their students, and even finding ways of leveraging the changes to help the teaching and learning process (for example, Jia, et al. (2020) used a variation of the flipped classroom model to improve student engagement). Crucially, these systems must be easy-to-use, and give control to the participants over what they choose to share. As mentioned previously, as well as privacy concerns, students have major concerns about judgement by their peers (Rajab and Soheib, 2021), so the systems must (both technically and procedurally) allow students to maintain the level of privacy that they desire. The outcomes of the *WebCam Usage Student Survey* (*WUSS*) address issues related to WebCam usage in general, as well as particularly in the case of e-Proctoring systems. The students' perspectives were varied and nuanced, and may indicate that the students are aware of the challenges of delivering educational content, and have been willing to forego aspects of their privacy for the sake of continuing their educational journey.

In the case of e-Proctoring systems, the key concerns relate to the potential lack of human agency in these systems, for example, if the systems are logging students out of an examination because of extraneous visual or audio inputs. However, it is worth noting that many of these systems do not take independent action, but rather notify a human proctor of suspected malfeasance, and the human must decide whether or not to take action. In fact, many of the concerns around these systems are as a result of the fact that they had to be rushed into service for such a wide variety of assessment processes in such a short period of time. As mentioned previously De Santis, et al. (2020) found that students who are knowledgeable about e-Proctoring systems are significantly more confident with their use in assessment, therefore it may the case that student anxiety about the use of these systems could abate if they are given more training on these systems, and more training on how they work. Additionally, it is important that teachers fully understand how these systems work so that they can instill confidence in their students.

It is hoped that discussions like these can serve as a reminder that all participants in the educational process have both rights and responsibilities in terms of their own privacy, and the privacy of others.

[14] https://www.forbes.com/sites/dianemehta/2013/04/26/new-survey-suggests-millennials-have-no-idea-what-privacy-means/?sh=20666b3229e2

CONCLUSIONS

This paper outlines an exploration of issues related to the use of webcams in an educational context, focusing in particular on some of the ethical considerations that have been exacerbated by the COVID-19 global pandemic. A review of some key literature is presented, focusing on some of those key ethical concerns, as well as presenting state-of-the-art research on the use of webcams since the onset of COVID-19. Following this the development of a survey to begin to capture the student voice in this discussion is presented, and the results of that survey are presented. The key outcome of the survey is that different students have different perspectives on these issues, so we are not seeing simplistic, binary, polarised thinking from students; the students who are most aware of the technological pitfalls of these systems, computer science students, understand that this is a nuanced issue with boons and banes, and therefore, to help educational organisations and individuals understand some of the challenges associated with the use of both WebCams and e-Proctoring systems, a discussion has presented, based on this work. The next step in this research is to create two sets of guidelines, one on webcam usage, and one on guidelines for e-Proctoring systems.

ACKNOWLEDGEMENTS

The authors of this paper and the participants of the Ethics4EU project gratefully acknowledge the support of the Erasmus+ programme of the European Union. The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein. The authors also acknowledge the invaluable pedagogical and technological advice provided by Dr. Ciaran O'Leary (Technological University of Dublin) in the preparation of this publication.

KEYWORDS: Digital Ethics, Privacy, e-Proctoring, Webcams.

REFERENCES

ACM, 2020, Association for Computing Machinery, *Statement On Principles And Prerequisites For The Development, Evaluation And Use Of Unbiased Facial Recognition Technologies*, Available at: https://www.acm.org/binaries/content/assets/public-policy/ustpc-facial-recognition-tech-statement.pdf

Adnan, M. and Anwar, K. (2020) "Online Learning amid the COVID-19 Pandemic: Students' Perspectives" *Journal of Pedagogical Sociology and Psychology*, 2(1), pp.45-51.

Allen, A.L., 2012. "What must we Hide: The Ethics of Privacy and the Ethos of Disclosure", *Thomas L. Rev.*, 25, p.1.

Andrejevic, M. and Selwyn, N. (2020) "Facial Recognition Technology in Schools: Critical Questions and Concerns", *Learning, Media and Technology*, 45(2), pp.115-128.

Arendt, Hannah (1998) "The Public and the Private Realm" In *The Human Condition*, pp. 182–230. Chicago, IL: University Of Chicago Press.

Aucejo, E.M., French, J., Araya, M.P.U. and Zafar, B. (2020) "The Impact of COVID-19 on Student Experiences and Expectations: Evidence from a Survey", *Journal of Public Economics*, 191, p.104271.

Chawki, M. and el Shazly, Y. (2013) “Online Sexual Harassment: Issues and Solutions”, *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, 4, p.71.

Chignell, M.H., Quan-Haase, A. and Gwizdka, J. (2003) The Privacy Attitudes Questionnaire (PAQ): Initial Development and Validation, In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 47, No. 11, pp. 1326-1330). Sage CA: Los Angeles, CA: SAGE Publications.

Coghlan, S., Miller, T. and Paterson, J. (2020) “Good Proctor or "Big Brother"? AI Ethics and Online Exam Supervision Technologies”. *arXiv preprint arXiv:2011.07647*.

Deleuze, Gilles (1992) “Postscript on the Societies of Control”, *October*, Vol. 59, pp. 3-7.

De Santis, A., Bellini, C., Sannicandro, K., Minerva, T. (2020) “Students’ Perception on E-Proctoring System for Online Assessment” In *EDEN 2020 Conference Proceedings*, No. 1, pp. 161-168.

Eaton, S.E. and Turner, K.L. (2020) “Exploring Academic Integrity and Mental Health during COVID-19: Rapid Review”, *Journal of Contemporary Education, Theory & Research*, 4(2), pp.35-41.

Foucault, Michel (1977) “Discipline and Punish: the Birth of the Prison”, New York: Random House.

González-González, C.S., Infante-Moro, A. and Infante-Moro, J.C. (2020) “Implementation of E-proctoring in Online Teaching: A Study About Motivational Factors”. *Sustainability*, 12(8), p.3488.

Habermas, Jürgen (1991) *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. Cambridge, MA: The MIT Press.

Jia, C., Hew, K. F., Bai, S., & Huang, W. (2020). “Adaptation of a Conventional Flipped Course to an Online Flipped Format during the COVID-19 Pandemic: Student Learning Performance and Engagement”, *Journal of Research on Technology in Education*, 1-21.

Kezer, M., Sevi, B., Cemalcilar, Z., Baruh, L. (2016) “Age differences in privacy attitudes, literacy and privacy management on Facebook”, *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 10(1).

Kshetri, N. (2020) “Remote education is rife with threats to student privacy”, The Conversation, November 6th, 2020. Full Content Available at: https://theconversation.com/remote-education-is-rife-with-threats-to-student-privacy-148955

Langenfeld, T. (2020) "Internet-Based Proctored Assessment: Security and Fairness Issues", *Educational Measurement: Issues and Practice*, 39(3), pp. 24-27.

Murtonen, M., Gruber, H. and Lehtinen, E. (2017) “The Return of Behaviourist Epistemology: A Review of Learning Outcomes Studies”, *Educational Research Review*, 22, pp.114-128.

O’Sullivan, D., Gordon, D. (2020) “Check Your Tech – Considering the Provenance of Data Used to Build Digital Products and Services: Case Studies and an Ethical CheckSheet”, IFIP WG 9.4 European Conference on the Social Implications of Computers in Developing Countries, 10th–11th June 2020, Salford, UK.

Rajab, M. H., Soheib, M. (2021) “Privacy Concerns Over the Use of Webcams in Online Medical Education During the COVID-19 Pandemic”, *Cureus*, 13(2).

Regehr, C. and McCahan, S. (2020) “Maintaining Academic Continuity in the Midst of COVID-19”, *Journal of Business Continuity & Emergency Planning*, 14(2), pp.110-121.

Remtulla, R. (2020) “The Present and Future Applications of Technology in Adapting Medical Education amidst the COVID-19 Pandemic”, *JMIR Medical Education*, 6(2), p.e20190.

Tufekci, Z. (2017) *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press.

Umawang, J. (2019) "Labs Survey Finds Privacy Concerns, Distrust of Social Media Rampant with all Age Groups", Date Accessed: 10/5/2021, Full Content Available at: https://blog.malwarebytes.com/security-world/2019/03/labs-survey-finds-privacy-concerns-distrust-of-social-media-rampantwith-all-age-groups/

Vatcha, A. (2020) "Workplace Surveillance Outside the Workplace: An Analysis of E-Monitoring Remote Employees" *iSCHANNEL*, p.4.

Zeide, E. and Nissenbaum, H. (2018) "Learner Privacy in MOOCs and Virtual Education". *Theory and Research in Education*, 16(3), pp.280-307.

Zuboff, S. (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Profile Books.

4. Ethics of Emerging Technologies

THE EVOLUTION OF PAYMENT METHODS IN MEXICO. ARE THEY ETHICALLY DISRUPTIVE TECHNOLOGIES?

Pedro I. González Ramírez, Juan Carlos Yáñez Luna

Universidad Autónoma de San Luis Potosí (México)

pedro.gonzalez@uaslp.mx; jcyl@uaslp.mx

ABSTRACT

The method of payment refers to the authorized mechanism by the financial system for performing the transaction of goods and services. The most instrument used in Mexico is cash, so in this study is identified that there is a lack in the acceptance and use of the other payment instruments in Mexico (for example, digital instruments). This study aims to complement applied research in the area of acceptance of technologies and the disruption and ethical implications. In addition. Outcomes for this study will expect that the application of the model described will confirm a positive factor regarding the acceptance of account holders of digital payment methods.

INTRODUCTION

A payment method (PM) is an asset that is used as money. In an economy, the PM is the basic tool for interchanging or acquiring goods and services by transferring monetary assets. There are two kinds of PM: Banknotes and Coins. The process for monetary assets transferring is simple, first the available balance of the individual and second, the mechanism to make the financial transaction (hard cash). However, the mechanisms for paying methods have been evolved, such as debit or credit cards, cashier or personal checks, electronic transfer, mobile apps, etc.

For the proper functioning of payment systems, it is essential to have reliable, efficient, and secure financial instruments. These instruments must provide certainty to economic agents that their financial transactions will be completely secure. Many countries have opted for technological instruments to improve their efficiency. The trend in this context has been the adoption of a complicated telecommunications infrastructure and specialized software. In this sense, a problem can be observed: on the one hand, developing and emerging countries have difficulties in their implementation, on the other hand, the vague or null financial education of these countries generates a lot of mistrust among economic agents.

For example, credit card payments are a payment method associated with a line of credit granted by a bank. This bank is called the issuing bank (IB). The IB undertakes to pay for transactions carried out in establishments authorized to use its cards. On the other hand, suppliers of goods and services in accepting these methods of payment must establish a relationship with a bank. This bank is called the Acquiring Bank (AB). The AB must install Point of Sale (POS) Terminals to process possible transactions. To carry out these transactions, the participation of other financial entities is required: The processing companies that oversee the communication services between the banks and the Card Payment Associations (CPA). The CPAs will be responsible for establishing the operational and financial rules between the IB and the AB. As can be seen in the example, the transaction process seems simple, however, it requires a complicated infrastructure. This could cause distrust in consumers, and therefore not adopt the mechanisms.

According to the report "Financial Inclusion Panorama 2020", cash transactions have increased during the period from December 2010 to December 2018. The balance of banknotes and coins has experienced an average annual growth rate of 7.1% as a percentage of GDP. The largest number of transfers was carried out by internet banking (52%). POS terminals registered 975.5 million operations, while ATMs 517.4 million operations, in addition, transfers registered 350.4 million operations, followed by electronic commerce with 119.8 million operations and checks 50.5 million operations. Although a growing dynamism can be observed in the use of various payment methods, Mexico is still below the same dynamism as countries with similar per capita income (Del Río Chivardi et al., 2020).

According to the International Payment Bank, Mexico has 101 POS for every 10,000 adults and 29 POS transactions per adult per year, while countries like Brazil have 317 POS for every 10,000 adults and 79 transactions per adult per year. Turkey, for its part, has 394 POS for every 10,000 adults and 70 POS transactions per adult per year.

The paragraph above shows the continuing popularity of using cash to transact. As indicated in Hancock & Humphrey (1998), cash has characteristics that make it the most traditional method of payment in the world, they are Practicality, Divisibility, and Acceptance. When a minor transaction is made in cash the resources can be used immediately. However, when a transaction is of a larger amount, the use of cash represents certain disadvantages, such as an increase in the probability of illegal attacks. There is another implicit cost of using cash that can have negative effects on economic activity. This cost refers to the loss or deceleration of transactions, on the one hand, due to the lack of cash in the establishments and on the other hand the non-acceptance of these payment methods by them.

In recent times, technological advances (supported by generational gaps) have intensified the use of digital financial services in the world. This is some way to forces banking institutions to exploit the internal market through transfer systems and internet-based payment methods.

According to Asociación de Internet MX (2019), the Study on the Financial Services of Internet Users in Mexico 2019, 75% of Internet users has some financial service. The most used are credit and debit cards (mostly to make transactions online). The report also indicates that only 3% of people have downloaded and used financial systems (mobile apps). In turn, it points out that the main barriers for non-users are lack of liquidity, as well as the perception of not needing any service and lack of trust towards financial institutions. Users of financial services want to get from their financial service lower commissions, complete security, and the fulfilment of promises.

The previous scheme allows us to observe that the acceptance of other methods of payment (technological) can have important effects on the volume of transactions carried out in an economy. In this sense, it would be relevant to identify the acceptance factors that influence the end-users of these technologies. However, some factors prevent this correct acceptance. According to Tanwar et al. (2018), most users have concerns about privacy and their rights when using financial technologies. These concerns point to the possibility that malicious people carry out fraud, identity theft, civil liberties, among others. In this context, Hajebrahimi et al. (2018) and Noordin et al. (2018) pointed out that disruptive technologies built new markets and new cost networks while perturbing the existing markets with disruption. Windell & Kroeze (2009) shows that disruption could have negative economic effects in organizations due to the challenges and changes in a global market. In this study is assumed that technologies could be a real inconvenience for the strategic positioning for banking institutions due to the increase in people's mistrust of them.

LITERATURE REVIEW

The concept of globalization has been consolidated in recent decades through the evolution of information technologies. It has been observed that national markets adopt the trend of internationalization by expanding their borders and consolidating supply chains (León-Peña, 2008). In this sense, companies that compete in global markets have had to invest in technological infrastructure to compete in them. This investment is part of the consolidation of business strategies to be able to survive in the markets (Porter, 2003). The needs of companies in general terms will depend on the changes in the environment where they are competing. Thus, access to fast and timely information allows companies to develop information banks, not only about their customers but also about what happens in their environment, thus allowing them to generate market strategies and obtain competitive advantages from it.

In economic terms, the topic of technological disruption refers to the first concepts on innovation and its management described by Schumpeter, referring to the creative destruction of things or technology-push, generating new adoptions in the market. Also, the disruption will depend on the context where it is applied, that is; refers to the changes that it can have concerning time, innovation can arise through partial and continuous changes until considering radical changes in the product, in the production systems, or the strategic decisions of the company. As already mentioned, the term refers to a radical transformation. Disruption focuses on a transformation in the company about its business model and technological advance and its speed of change (Bower & Christensen, 1995).

A more specific concept is exposed in Mendoza-Tello et al. (2019). The authors point out that disruptive innovation is defined as a metric that directly affects certain components such as performance, consumer expectations, or market behaviour in the face of radical innovation. Zubizarreta et al. (2021) point out that disruptive innovation alters the characteristics of products and services in the market, providing them with greater value.

The previous concepts are not far from the initial line of the concept. In Christensen (2013) the effect of the entry of technologies for the success of the company is studied. In this text, the author describes some principles of disruptive technology that have an implication with the company and its success in adapting to these changes. The first principle explains the business-consumer relationship and the degree to which markets demand certain needs and are disparate with current advances in technology. This can cause a lack of interest in the consumer at a certain time, however, that need may develop in the future. The second principle is related to the reflection of innovation management with the allocation of resources. Christensen points out that this difficulty is influenced by the difference between the experience of those who implement the decisions and those who make the decisions, so it will be difficult for top managers to allocate resources to develop disruptive technologies. The third principle focuses on observing disruptive technologies as part of marketing planning. This means that entrepreneurs must know the market or, failing that, they must find a market that values the disruptive characteristics of the product.

The fourth principle in Christensen's general scheme indicates that the production capacities of companies are focused on their value networks and this allows finding new market niches in which to position certain products under certain characteristics in the environment. These activities are reflected in its behaviour around its profitability. The fifth principle adds the value of information and the risk that it carries. The author points out that access to information is not necessarily important to make important investments around disruptive technologies, but rather, information must be generated through quick access, establishing economic and flexible strategies in the market and the product. The sixth principle suggests a determination about the position that the company has regarding the technology that it wants to enter the market (disruptive or sustainable). Finally, the

author points out that there are entry and mobility barriers. This topic refers in a generalized way to the entrepreneur's mentality in investors and the difficulties to adjust the market changes in their business model, which constitute entry and mobility barriers applicable in the market.

The previous points offer an overview of what the topic of disruptive technologies offers in a market environment in general. It is assumed in economic terms that continuous technological changes can lead to technological disruption and in turn to radical changes in the markets or establishment of business strategies, and this would have implications for the environment. This research work will focus on terms of observing the perception of consumers about the possible effects of technological disruption in banking institutions.

As already mentioned, technological advance has increased exponentially. Banking institutions have not been the exception in this context. Financial services require more than ever a constant integration of technological means, on the one hand, to compete in the market and on the other hand to offer better services to their clients (Vives, 2019). The integration of technologies oriented to the cloud machine learning will require strong investments in addition to a strategic change at the corporate level to position itself within the market as a pioneer. These changes are often of particular interest to the leaders of these institutions. A study carried out by Forbes Staff (2020) indicates that from the period of health contingency worldwide, the use of financial services through electronic devices increased. This change suggests a process of disruption in the business strategies of financial groups, that is, orienting their model to technological and digital strengthening, not only to end-user applications but also to the technology that supports the service.

In this sense, those responsible for financial institutions must consider studying the market in which they are competing. These studies should focus on the way that service consumers adopt and adapt to new systems and what their main implications are. Carbo Valverde & Rodríguez Fernández (2019) point out that acceptance strategies should not only be based on aspects of market demand, but it is also important to consider aspects of supply; that is, designing proposals that allow satisfying the digital needs detected in consumers of financial services.

Cryptocurrencies have gained ground in the last decade as a method of payment. According to Mendoza-Tello et al. (2019), this type of unregulated method of payment can be considered a disruptive innovation due to its high technological use. The application of technological resources, especially in services, also leads to establishing security strategies and establishing trust in the consumer. In the case of cryptocurrencies, the authors point out three key points to consider: User trust, the usefulness of the instrument, and popularity.

Concerning the above and again following Christensen cited in (Gobble, 2016), he points out that disruptive innovations regularly originate under certain characteristics based on low-level support points, that is, low-end products or services. In this sense, it is very difficult for financial institutions to establish a low-end service. On the one hand, certainly, the disruption could focus on serving an unattended market niche, however, this can be considered partially disruptive, since consumers of financial services can access another method of payment other than digital. On the other hand, the disruption could focus in a better way on the quality of the service that is being provided to the consumer, the disruption strategies would have to focus on the characteristics indicated by Christensen. Faced with this scenario, this research work aims to develop a model that makes it possible to explain to decision-makers of financial services the degree of acceptance that consumers have towards their digital services. This leads us to question the following, given the substantial changes in technology worldwide, added to the adoption of these by institutions, can this evolution be considered as a disruptive action in the market?

Another important point to highlight in this type of research is to evaluate the degree of acceptance that consumers have of this technological revolution. The next section will roughly review the most widely used technology acceptance models to perform these measurements.

TECHNOLOGY ACCEPTANCE MODELS

As mentioned in the previous section, it was commented on the needs of consumers and the strategic bases that businesses must follow to satisfy them. In recent times, businesses and economies in the countries were affected by COVID19, however, financial institutions managed to establish strategies to enhance their digital services.

The above, in business terms, is directly related to consumer satisfaction. This action also increases trust and, in turn, loyalty towards the institution (Delgado-Ballester, 2003). Many studies focus on explaining the influence of certain variables (such as a trust) on consumer behaviour regarding the acceptance of a product or service. There are various quantitative methodologies in the academic literature that allow measuring the degree of acceptance through trust, however, the results that could be obtained would vary with the context where they are being applied, for example, individual or collective contexts, based on accumulated experiences, expected results, social or political environment, perceived risks, etc.

The evaluation of digital tools for financial services goes beyond familiarity as an enhancement of trust and the intention of use. The consumer's attitude towards the adoption of an innovation can be traced back to studies such as the Theory of Reasoned Action (TRA) (Fishbein & Ajzen, 1975). This theory is based on describing the attitudes of individuals and their association with key objects. As a result of these studies, new models have developed that attempt to explain the acceptance and intention to use technology specifically for information systems, the most common being Technology Acceptance Model (TAM) developed by (Davis, 1985). Davis suggests that both the attitude towards the use as well as the intention to use technology will depend on the perception of utility and the perception of ease of use of the technology. These perceptions in turn will be influenced by external factors with emphasis on technical and social aspects and factors that have moderating implications among the variables. Venkatesh, Morris, Davis, & Davis (2003) elaborate a unified theory that tries to explain the intention of adoption and use of technologies. This theory arises because of TAM and other models of technology acceptance. Like TAM, UTAUT suggests that four factors (expectation of effort, expectation of performance, social influence, and the facilitation of conditions) explain the behaviour towards the use and in turn the intention to use the technology.

Although it has been commented throughout this work that a disruption in financial services is not based one hundred percent on their technological structure if it can be said that they make use of technology to develop in the market, and this would lead to a disruption progressive. Therefore, it is essential that the institutions that work in this market develop strategies based on precise elements that allow explaining the degree of consumer acceptance. In this sense, sociological currents study and try to explain (from a micro-social perspective) the relationships between individuals and their behaviour within a social entity. Consumer behaviour is highly relevant, and its study allows us to identify new opportunities and market segmentation, taking into consideration that human beings by nature are a social entity and will always seek to link into social groups that are ad hoc to their needs or tastes. These links can influence the behaviour of the individual and their future decisions (Noor et al., 2013), for example in other types of products or services such as fashions in clothing, music, books, etc.

A factor mostly used by researchers in the study of consumer behaviour is social influence. Studies on this factor are very diverse, for example, Venkatesh, Morris, & Ackerman (2000) point out that “it refers to the perceived social pressure to influence behaviour or not”. In this case, the norms attached to society have a higher impact on the individual in a degree of "obedience" rather than by conviction. Initially, this factor was introduced in the acceptance models to measure the degree of the implication that co-workers or higher orders have in the personnel that will operate a certain system in a company. Subsequently, the construct was used not only to measure its involvement in a closed context but in a more extensive context that encompasses the consumer's social environment. In this way, studies of acceptance of technology applied to technological implants have been carried out using cognitive-affective-normative models (Pelegrín-Borondo et al., 2016) where the role of subjective norms has direct implications on the consumer at the time of deciding to use or adopting the innovation.

Regarding the digital aspects, it is common to find variations in the acceptance models. These variations are adjustments that researchers make to try to get closer to the reality of the environment. Oliveira et al. (2014) carried out a study to know the degree of acceptance of Mobile Banking through a combination of main constructs of three acceptance models UTAUT, Initial Trust Model (ITM), and Task-Technology-Fit (TTF). They found that this combination explained some determinants in consumer behaviour regarding their intention to accept. These determinants explain the initial confidence, the expectation of performance, and the characteristics of the technology. Another study (Warsame & Ireri, 2018) points out the use of some moderating variables in the acceptance models. Some of them are directly related to normative elements such as religious beliefs, trust, and social influences. The result of the study showed an influence of these variables on the intention to use microfinance and mobile banking in everyday use environments. Similarly, Baptista & Oliveira (2015) pointed out that habits and culture were the variables with the greatest impact in reducing uncertainty and explaining the acceptance of mobile banking.

In previous paragraphs, it was pointed out that the perceived quality of a product or service is an element that the consumer values highly and can be a fundamental factor in reducing the perceived risk. The study of perceived risk is found in Jacoby & Kaplan (1972). The authors identify five variants of perceived risk: financial, performance, physical, psychological, and social risk. The behaviour of these variables is usually independent, each one concerning the others, that is, if the perception of risk increases in one, the others may remain with the same value, however, variables may also present increases or decreases in the values. The psychological and social risk factors respond directly to the general concept of subjective norms and could have implications in ethics. Several studies (Jaradat et al., 2018; Lee, 2009; Senyo & Osabutey, 2020) have focused on explaining the perceived risk, however, the results differ in each study. While some authors point out the importance of the perceived financial risk around the intention of use, other authors point out that its effects are not considered. Despite the above, the importance of evaluating these variables in mobile banking, inclusion, and financial services environments, in general, are considered.

THE MODEL

The ordered PROBIT model is used to estimate relationships between a dependent, ordinal, and categorical variable and a set of independent variables. Examples of ordinal and categorical variables are values on statements such as: "completely disagree", "disagree", "somewhat agree" or "completely agree". If the categorical variable has only two results, a traditional PROBIT model is estimated; if the variable has more than two results and they cannot be ordered, a multinomial model is estimated.

In the ordered PROBIT model, the underlying value is estimated as a linear function of the independent variables and a set of cut-off points. The probability of observing the outcome i corresponds to the probability that the estimated linear function, plus random error, is within the range of cut-off points estimated for the outcome.

$$P_r\big(outcome_j = i\big) = P_r(k_{i-1} < \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_m x_{mj} + u_j \leq k_i) \quad (1)$$

Where u_j is assumed to be a normal distribution. The model is a generalization of the probit model for two outcomes. In ordinal models, the β_i allows us to identify the direction of change in the probability in the presence of a change in the variable x_i, only that it is not as simple as the traditional model because now more than two outcomes must be calculated, in other words:

$$\frac{\partial P_r(y = k_1|x)}{\partial x_i}, \frac{\partial P_r(y = k_2|x)}{\partial x_i}, \ldots, \frac{\partial P_r(y = k_n|x)}{\partial x_i} \quad (2)$$

Considering that these changes depend on the number of categories k_i, which are established in the dependent variable y.

The mentioned before is applicable in technology acceptance models, where the intention to adopt or use technology and the response, which is typically given in an ordinal and categorical form according to a Likert scale, is intended to be explained; the application of the ordered Probit model is more than relevant since it allows measuring how each of the constructs x_i can affect the probability in the intention to use such technology.

Therefore, the model can be stated as follows: the dependent variable y is defined as how much a person agrees to adopt a new technology, where this variable can take the following values: "completely disagree=1", "disagree=2", "slightly agree=3" and "completely agree=4". The ordered probit model allows us to calculate, based on the explicative variables and the error component, what is the probability that a person is willing to accept a technology among a cut-off of values defined for the variable y, that is:

$$P_r(somewhat\ agree < \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_m x_{mj} + u_j \leq completely\ agree) \quad (3)$$

Consequently, this methodology complements the analysis of the TAM and UTAUT models, since it quantifies how changes in the explanatory variables can affect the probability of acceptance of the technology, as well as measuring which of these variables has the greatest impact on the probability of acceptance, and not only analyzes the relationship between the explanatory variables and the intention to adopt.

DISCUSSION

This study aims to complement applied research in the area of acceptance of technologies or the context of innovation. In addition, the area of acceptance of payment methods (traditional and digital) in Mexico has not yet been exploited, so this work aims to be a pioneering study in this emerging area of financial systems. Although this work is in process, it is expected that the application of the model described in this study will confirm a positive factor regarding the acceptance of account holders of digital payment methods.

Some expected contributions will be discussed in this study. In the first instance, technology acceptance models are regularly based on exploratory methodologies and not so much on confirmatory ones (Chin & Dibbern, 2010). The methodology used in this study brings the researcher closer to a more probable reality in the environment, that is, a verifiable analysis (García de los Salmones et al., 2006). This research work will try to take as a confirmatory basis the main factors of the models of acceptance and use of technology, for example, the perceived utility and ease of use (Davis, 1985) complemented with other factors such as social influence, effort expectations and performance and enabling conditions (Venkatesh et al., 2003). The study will seek to demonstrate that there is a high probability that consumers of financial services accept the use of digital technology concerning the results of (Oliveira et al., 2014; Zhou et al., 2010) where direct implications were found between the effort and performance expectations factors in the acceptance of digital banking services.

In a second instance, this study will integrate the risk factor as a predictor of social influence and with disruptive implications in the organization (Windell & Kroeze, 2009), this means that the more experience the consumer of financial services has, trust and confidence will increase. consumer uncertainty will decrease (Sanchez-Franco, 2009). Financial risks comprise several areas (Jacoby & Kaplan, 1972), however, this study will discuss the approach that financial and social risk areas have towards the main factors of technology acceptance. Due to the nature of the construct, the measurement of this factor is expected to be negative due to uncertainty, lack of information, and consumer experience.

The discussion above shows several factors that influence the acceptance of technology. These factors have been studied over time in various areas. This study proposes a new consideration for the main factors that influence the acceptance of technology by including the factors that will influence the social and ethical perspective, as well as consumer trust and safety. The equations above show the theoretical framework for forecasting the acceptance of technologies, which for the case study will be digital payment methods. The equations were modelled based on the ordered PROBIT methodology. This methodology will provide the base structure for the elaboration of the research hypotheses.

Finally, the results of this research will have implications for the theoretical basis of technology acceptance models. There are various points of view in the area, however, this study aims to show the potential of social influence and its possible diversifications, as well as factors that can help to better explain the social construct. The development of these models will also meet the practical needs of both researchers and decision-makers in banking institutions regarding a better understanding of consumer behaviour and its constant evolutions.

KEYWORDS: Payment methods, Banking, Disruptive Technologies, Financial market.

REFERENCES

Asociación de Internet MX. (2019). *Estudio sobre Comercio Electrónico en México 2019. Décima tercera entrega*.

Baptista, G., & Oliveira, T. (2015). Understanding mobile banking: The unified theory of acceptance and use of technology combined with cultural moderators. *Computers in Human Behavior*, *50*, 418-430. https://doi.org/10.1016/j.chb.2015.04.024

Bower, J. L., & Christensen, C. M. (1995). Disruptive Technologies: Catching the Wave. *Harvard Business Review*, *February*, 43-53. https://hbr.org/1995/01/disruptive-technologies-catching-the-wave

Carbo Valverde, S., & Rodríguez Fernández, F. (2019). Patrones de acceso a la banca digital: aproximaciones tradicionales, aprendizaje automático y neuroeconomía. *Papeles de Economía Española*, *162*, 14-26.

Chin, W. W., & Dibbern, J. (2010). PLS PathModeling: From Foundations to Recent Developments and Open Issues for Model Assessment and Improvement. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of Partial Least Squares. Concepts, Methods and Applications* (Issue July, pp. 171-193). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-32827-8

Christensen, C. M. (2013). *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business Review Press.

Davis, F. D. (1985). *A Technology Acceptance Model for Epirically Testing New End-User Information Systems: Theory and Results*. Massachusetts Institute of Technology.

Del Río Chivardi, M. A., Castro Solares, C. E., Hernández Godínez, J., & Cano Vallejo, S. R. (2020). *Panorama Anual de Inclusión Financiera*.

Delgado-Ballester, E. (2003). Development and validation of a brand trust scale. *International Journal of Market Research*, *45*(1), 35-54. https://doi.org/10.1088/1751-8113/44/8/085201

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research.* Addison-Wesley.

Forbes Staff. (2020). *Los mejores bancos del mundo 2020: el futuro de la banca es digital - Forbes Colombia*. Forbes. Negocios. https://forbes.co/2020/06/08/negocios/los-mejores-bancos-del-mundo-2020-el-futuro-de-la-banca-es-digital/

García de los Salmones, M. del M., Herrero Crespo, Á., & Rodríguez del Bosque, I. (2006). Actuación comercial, imagen y lealtad: aplicación al sector B2B de acceso a redes de telecomunicaciones. *INNOVAR. Revista de Ciencias Administrativas y Sociales*, *16*(27), 101-116. http://www.redalyc.org/articulo.oa?id=81802708%0ACómo

Gobble, M. A. M. (2016). Defining disruptive innovation. *Research Technology Management*, *59*(4), 66-71. https://doi.org/10.1080/08956308.2016.1185347

Hajebrahimi, A., Kamwa, I., & Huneault, M. (2018). A novel approach for plug-in electric vehicle planning and electricity load management in presence of a clean disruptive technology. *Energy*, *158*, 975-985. https://doi.org/10.1016/j.energy.2018.06.085

Hancock, D., & Humphrey, D. B. (1998). Payment transactions, instruments, and systems: A survey. *Journal of Banking & Finance*, *21*, 1573-1624.

Jacoby, J., & Kaplan, L. B. (1972). The components of perceived risk. In M. Venkatesan (Ed.), *SV - Proceedings of the Third Annual Conference of the Association for Consumer Research* (pp. 382-393). Association for Consumer Research.

Jaradat, M.-I. R. M., Moustafa, A. A., & Al-Mashaqba, A. M. (2018). Exploring perceived risk, perceived trust, perceived quality and the innovative characteristics in the adoption of smart government services in Jordan. *International Journal of Mobile Communications*, *16*(4), 399-439. https://doi.org/10.1504/IJMC.2018.092669

Lee, M.-C. (2009). Factors influencing the adoption of internet banking: An integration of TAM and TPB with perceived risk and perceived benefit. *Electronic Commerce Research and Applications*, *8*(3), 130-141. https://doi.org/10.1016/j.elerap.2008.11.006

León-Peña, J. R. (2008). E-Business And The Supply Chain Management. *Business Intelligence Journal*, *1*(1), 77-90.

Mendoza-Tello, J. C., Mora, H., Pujol-López, F. A., & Lytras, M. D. (2019). Disruptive innovation of cryptocurrencies in consumer acceptance and trust. *Information Systems and E-Business Management*, *17*(2-4), 195-222. https://doi.org/10.1007/s10257-019-00415-w

Noor, M. N. M., Sreenivasan, J., & Ismail, H. (2013). Malaysian Consumers Attitude towards Mobile Advertising, the Role of Permission and Its Impact on Purchase Intention: A Structural Equation Modeling Approach. *Asian Social Science*, *9*(5), 135-154. https://doi.org/10.5539/ass.v9n5p135

Noordin, M. F., Othman, R., & Rassa, A. H. R. (2018). Social Media and Knowledge Management Disruptive Technology. *Proceedings of Knowledge Management International Conference (Kmice) 2018*, *July*, 6-11.

Oliveira, T., Faria, M., Thomas, M. A., & Popovic, A. (2014). Extending the understanding of mobile banking adoption: When UTAUT meets TTF and ITM. *INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT*, *34*(5), 689-703. https://doi.org/10.1016/j.ijinfomgt.2014.06.004

Pelegrín-Borondo, J., Reinares-Lara, E., Olarte-Pascual, C., & Garcia-Sierra, M. (2016). Assessing the Moderating Effect of the End User in Consumer Behavior: The Acceptance of Technological Implants to Increase Innate Human Capacities. *Frontiers in Psychology*, *7*(February), 132. https://doi.org/10.3389/fpsyg.2016.00132

Porter, M. E. (2003). Strategy and the Internet. *Harvard Business Review*, 01-20. https://hbr.org/2001/03/strategy-and-the-internet

Sanchez-Franco, M. J. (2009). The Moderating Effects of Involvement on the Relationships Between Satisfaction, Trust and Commitment in e-Banking. *Journal of Interactive Marketing*, *23*(3), 247-258. https://doi.org/10.1016/j.intmar.2009.04.007

Senyo, P. K., & Osabutey, E. L. C. (2020). Unearthing antecedents to financial inclusion through FinTech innovations. *TECHNOVATION*, *98*. https://doi.org/10.1016/j.technovation.2020.102155

Tanwar, S., Tyagi, S., Kumar, N., & Obaidat, M. S. (2019). Ethical, Legal, and Social Implications of Biometric Technologies. In M. Obaidat, I. Traore, & I. Woungang (Eds.), *Biometric-Based Physical and Cybersecurity Systems* (pp. 535-569). Springer International Publishing. https://doi.org/10.1007/978-3-319-98734-7_21

Venkatesh, V., Morris, M. G., & Ackerman, P. L. (2000). A Longitudinal Field Investigation of Gender Differences in Individual Technology Adoption Decision-Making Processes. *Organizational Behavior and Human Decision Processes*, *83*(1), 33-60. https://doi.org/10.1006/obhd.2000.2896

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425-478. https://doi.org/10.2307/30036540

Vives, X. (2019). La banca frente a la disrupción digital. *Papeles de Economía Española*, *162*, 2-13. https://www.funcas.es/wp-content/uploads/Migracion/Articulos/FUNCAS_PEE/162art02.pdf

Warsame, M. H., & Ireri, E. M. (2018). Moderation effect on mobile microfinance services in Kenya: An extended UTAUT model. *JOURNAL OF BEHAVIORAL AND EXPERIMENTAL FINANCE*, *18*, 67-75. https://doi.org/10.1016/j.jbef.2018.01.008

Windell, A. C., & Kroeze, J. H. (2009). The use of disruptive information technologies for competitive advantage in the banking sector of South Africa. *Creating Global Economies through Innovation and*

Knowledge Management Theory and Practice - Proceedings of the 12th International Business Information Management Association Conference, IBIMA 2009, *1-3*(May), 1404-1411.

Zhou, T., Lu, Y., & Wang, B. (2010). Integrating TTF and UTAUT to explain mobile banking user adoption. *COMPUTERS IN HUMAN BEHAVIOR*, *26*(4), 760-767. https://doi.org/10.1016/j.chb.2010.01.013

Zubizarreta, M., Ganzarain, J., Cuadrado, J., & Lizarralde, R. (2021). Evaluating disruptive innovation project management capabilities. *Sustainability (Switzerland)*, *13*(1), 1-22. https://doi.org/10.3390/su13010001

SOCIAL MEDIA USER EMOTIONS DURING COVID19

Jan Strohschein, Ana María Lara-Palma, Heide Faeskorn-Woyke

TH Köln (Germany), Universidad de Burgos (Spain), TH Köln (Germany)

jan.strohschein@th-koeln.de; amlara@ubu.es; heide.faeskorn-woyke@th-koeln.de

ABSTRACT

Social networks are everywhere and a large part of users even frequents more than one platform (Pew Research Center, 2018). "Due to a constant presence in the lives of their users, social networks have a decidedly strong social impact" (Statista, 2019). However, several studies also suggest that social media usage is not beneficial for users health with symptoms ranging from sleep deprivation to anxiety and depression (Hogue & Mills, 2019; Hunt et al., 2018; Levenson et al., 2016).

This work uses a machine learning approach to study the emotions of a large group of social media users on Twitter during the Covid19 pandemic and compares the results to our previous research that evaluated 10 million tweets from 5000 users between 2015 - 2019.

It is possible to extract emotions of social media users from the text of their status updates as shown by Colneric and Demsar, and Tasoulis et al. (Colneric & Demsar, 2018; Tasoulis et al., 2018). This analysis is based on the work of Colneric and Demsar, who were kind enough to publish the resulting machine learning model. They utilized neural networks to generate a model that is able to detect emotions in English language. Neural networks are a supervised machine learning method and therefore the data needs annotations for the algorithm to learn from. As the authors learned on a massive dataset of 73 billion tweets it was infeasible to manually annotate the dataset. The authors exploited hashtags as annotations, an approach that was successfully used in several other natural language processing studies for sentiment classification (Go et al., 2009; Kouloumpis et al., 2011; Nodarakis et al., 2016), detecting sarcasm (Bamman & Smith, 2015; González-Ibáñez et al., 2011), studying personality traits (Plank & Hovy, 2015) and classifying emotions (Mohammad & Kiritchenko, 2015). As hashtags are selected by the author of a tweet they work well as indicators of their emotions.

Emotions can be modelled in a multitude of ways and popular emotion classification schemes were created by Paul Ekman, Robert Plutchik and Douglas McNair along with Maurice Lorr and Leo Droppleman (Ekman, 1999; McNair et al., 1971; Plutchik, 1982). The classification for this analysis is done with Ekmans scheme of basic emotions as it covers fear, disgust and anger, which have been previously identified as the most impactful emotions caused by the use of social media and should be investigated further.

In our previous work users have been grouped based on the number of status updates they publish and the amount of followers they have. Grouping users was more effective and showed more distinct results when based on the number of followers a user has. The prevalent expressed emotions on twitter from 2015-2019 were joy and surprise. Over the observed period of time, from 2015 to 2019, the values for joy remained consistent, while an increase in anger, disgust and fear could be verified for all user groups. It was noticeable that twitter users with the least amount of followers (<25%) expressed anger and fear most strongly. Even though it was expected that the positive emotions declined and the negative emotions increased during Covid19 the exact opposite happened. A detailed monthly analysis of the data suggests that the United States election had a big influence on the results.

INTRODUCTION

Most social media users frequent more than one platform (Pew Research Center, 2018). "Due to a constant presence in the lives of their users, social networks have a decidedly strong social impact" (Statista, 2019). "The blurring between offline and virtual life as well as the concept of digital identity and online social interactions are some of the aspects that have emerged in recent discussions. Approximately 2 billion, mostly young, internet users are using social networks and these figures are still expected to grow as mobile social network usage increasingly gains traction" (Statista, 2017, Statista, 2016, Pew Research Center, 2019). This paper builds upon earlier work that analyzed Twitter status updates from 2015-2019 (Strohschein et al., 2019). It investigates the development of emotions for different user groups on Twitter during the Covid19 pandemic utilizing machine learning and natural language processing techniques to create an automated approach.

Social Media Platforms

The usability of social media platforms has been tested in multiple fields being beneficial in a large number of indicators, such as learning (Hortigüela-Alcalá et al., 2019), inclusion, or socialization. However, unfortunately, for a percentage of the world's population, the technological mirror in which people look returns the image of sadness, fear, worry, and hopelessness. Many studies suggest that social media usage is not beneficial for user´s health with symptoms ranging from sleep deprivation to anxiety and depression (Levenson et al., 2016; Hunt et al., 2018; Hogue & Mills, 2019). Regarding these emotional consequences, Yoon, Kleinman, Mertz, and Brannick (Yoon et al., 2019) in their meta-analysis study on the correlation between social networks and symptoms of depression, highlight “Our results are consistent with the notion of ‘Facebook depression phenomenon’ and with the theoretical importance of social comparisons as an explanation”. But the effects of sleep deprivation and depression also persist during the workday and companies fear for their organizational productivity. Several groups of researchers studied the effects of social media on the productivity of students and workers alike. Brooks conducted research on students and supposes that being in the classroom can be analogous to being in a work environment as the students have to efficiently perform different tasks. He found that inefficiencies result from time spent on the interruption but also the time necessary to fully concentrate on the task again (Brooks, 2015). Lau as well as Flanigan and Babchuk also concluded that social media usage decreases motivation and hinders academic performance (Lau, 2017; Flanigan & Babchuk, 2015). Ali-Hassan, Nevo, and Wade studied the effects of social media in the workplace and found that social use of technology can have an indirect positive effect on job performance by building networks in the workplace and sharing knowledge but also discovered a direct negative impact on task routine performance when workers spent their time in social networks instead (Ali-Hassan et al., 2015). While research shows no clear results whether or not social media platforms hinder work performance, some companies don’t want to risk a loss in productivity and try to ban social media from the workplace (Gaudin, 2009).

Users and User groups

Social media users and digital natives have been subject to a lot of studies. Oblinger & Oblinger (Oblinger & Oblinger, 2005) characterized them as active experiential learners, proficient in multitasking, dependent on communication technology to access information and interacting with others. Kennedy, Judd, Delgarno and Waycott define digital natives based on several parameters, e.g., the number of devices they regularly use, their formal education, gender and age, to create the following user groups: power user, ordinary user, irregular user and basic user (Kennedy et al., 2010).

Unfortunately, those parameters are not easy to obtain in an automated approach. Twitter profiles do not contain any information about the formal education of a user. Specifying a location or a birth date is completely optional and even if a location is specified, it does not have to be in a standardized format so “Berlin” could mean the German capital or one of several cities with this name in the USA or Australia. Therefore, for this analysis users are grouped solely based on their number of followers, who want to read new status updates because this approach has been effective in the previous work.

Emotion Modeling

Emotions can be modeled in a multitude of ways and popular emotion classification schemes were created by Paul Ekman, Robert Plutchik and Douglas McNair along with Maurice Lorr and Leo Droppleman (Ekman, 1999; Plutchik, 1980; McNair et al., 1971). The classification for this analysis is done with Ekmans scheme of basic emotions as it covers fear, disgust and anger, currently the most researched emotions in association with social media networks according to the comprehensive literature review of our previous work. The six basic emotions are explained by Ekman and Cordaro (Ekman, Cordaro, 2011) in a later article as follows:

Anger: the response to interference with our pursuit of a goal we care about. Anger can also be triggered by someone attempting to harm us (physically or psychologically) or someone we care about. In addition to removing the obstacle or stopping the harm, anger often involves the wish to hurt the target.

Fear: the response to the threat of harm, physical or psychological. Fear activates impulses to freeze or flee. Often fear triggers anger.

Surprise: the response to a sudden unexpected event. It is the briefest emotion.

Sadness: the response to the loss of an object or person to which you are very attached. The prototypical experience is the death of a loved child, parent, or spouse. In sadness there is resignation, but in can turn into anguish in which there is agitation and protest over the loss and then return to sadness again.

Disgust: repulsion by the sight, smell, or taste of something; disgust may also be provoked by people whose actions are revolting or by ideas that are offensive.

Joy: feelings that are enjoyed, that are sought by the person. There are a number of quite different enjoyable emotions, each triggered by a different event, involving a different signal and likely behavior. The evidence is not as strong for all of these as it is for the emotions listed above.

Emotion Classification

It is possible to extract emotions of social media users from the text of their status updates as shown by Tasoulis et al. and Colneric and Demsar (Tasoulis et al., 2018; Colneric & Demsar, 2018). This analysis is based on the work of Colneric and Demsar, who utilized deep learning of neural networks to generate a model that is able to detect emotions in english language. Neural networks are a supervised machine learning method and therefore the data needs annotations for the algorithm to learn from. As the authors learned on a massive dataset of 73 billion tweets it was infeasible to manually annotate the dataset. The authors exploited hashtags as annotations, an approach that was successfully used in several other natural language processing studies for sentiment classification (Go, et al., 2009; Nodarakis, et al., 2016; Kouloumpis, et al., 2011), detecting sarcasm (Gonzalez-Ibanez, et al., 2011; Bamman & Smith, 2015), studying personality traits (Plank & Hovy, 2015) and classifying emotions

(Mohammad & Kiritchenko, 2015). As hashtags are selected by the author of a tweet they work well as indicators of their emotions. The machine learning algorithm analyses the given text and tries to derive the hashtag as target variable.

Hypothesis

The following hypotheses are used to investigate the development of user emotions on the Twitter social media platform during the COVID19 pandemic in comparison to the previous years:

H1: Joy will decline for users.

H2: Anger will increase for users.

H3: Disgust will increase for users.

H4: Fear will increase for users.

H5: Sadness will increase for users.

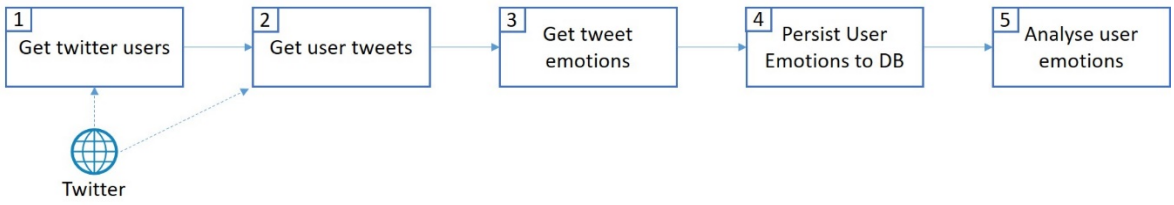
H6: Surprise will increase for users.

H7: It is possible to differentiate emotions between user groups, based on their number of followers.

METHODOLOGY

This paper uses an automated approach to study the emotions of a larger group of social media users over time. Colneric and Demsar published their resulting machine learning model on GitHub (Colneric & Demsar, 2018). This machine learning model is implemented in a real-time architecture to collect and analyze tweets and create a classification for the emotions expressed in the text. The CAAI architecture, developed by our workgroup, consists of virtualized processing, storage or analysis building blocks that communicate via messaging to create a highly modular data analytics pipeline, depicted in Figure 1.

Figure 1. Data collection and analysis pipeline.



In the first step trending Twitter topics are observed and active users, who write status updates on these topics, extracted. For every user their following users were extracted to increase the user sample-size. In a second step for every user the tweets were collected through the Twitter API. So called “retweets”, where a user quotes the status update of someone else, were not collected to analyse only tweets that the user wrote himself. The API imposes limits on the amount of requests per time interval and the number of status updates that can be downloaded for any given user. Therefore, only the last ~3.200 status updates can be retrieved for each user. Surprisingly this turned out to be

an important constraint, because for very active users this was just a fraction of their status update history.

The pre-trained model from Colneric and Demsar was used in the third step to analyze the status updates regarding expressed emotions. This step was run in parallel as the process was very time-consuming and the results were reconciled and persisted in a database (step 4) for further analysis (step 5). For the analysis the emotions of all users were evaluated over time.

RESULTS

For each user the last ~3200 tweets have been analysed, if the user wrote that many status updates. Over the course of several weeks roughly 11 million tweets from ~6000 users have been collected and analysed for the whole of 2020 and January to March of 2021.

Data Overview

The analysis is based on two related datasets, the users and their tweets with the associated emotional classification, that have been collected in separate steps but are joined to increase the available information. The two tables below describe the datasets and the available fields.

Table 1. User Features and descriptions.

User Feature	Description
user_id (integer)	Automatically generated identification number for every user.
user_name (string)	The user can choose the nickname to display.
user_location (string)	The user can specify his/her location.
account_created_at (timestamp)	The system records the day and time of account creation.
statuses_count (integer)	The amount of status updates a user has written, including retweets.
favorites_count (integer)	The number of Tweets this user has liked.
followers_count (integer)	The number of followers this account currently has.
friends_count (integer)	The number of users this account is following.
verified (boolean)	If the user´s identity has been verified by twitter the value is “True”, otherwise “False”.

Table 2. Tweet Analysis Features and Descriptions.

Tweet Analysis Feature	Description
status_id (integer)	Automatically generated identification number for every tweet.
user_id (integer)	Automatically generated identification number for every user.
status_created_at (timestamp)	The system records the day and time of tweet creation.
text (string)	The tweet text written by the user and analyzed for emotions.
retweet_count (integer)	Number of times this status update has been “retweeted”.
anger, disgust, fear, joy, sadness, surprise (float)	The calculated percentage value for a particular emotion in a users tweet.

An example for such a classification is shown below in Table 3, only the columns relevant for the classification are shown. The status update regards a sport event and the detected prevalent emotion is joy, followed by surprise.

Table 3. Tweet Analysis multi-class Classification Example.

Text	Anger	Disgust	Fear	Joy	Sadness	Surprise
'Anthony Davis makes his debut with the Hornets dropping 21 points and grabbing 7 rebounds.'	0.0109	0.0028	0.0425	0.7640	0.0551	0.1244

Distribution of tweets in the dataset per year

The previous analysis consisted mostly of status updates for 2018 and 2019 with few status updates for the earlier years. A possible reason is, that for very active users analysing 3200 tweets is just not enough. Another possibility may be that just active users are analysed and users from the earlier years stopped using the platform. The current evaluation contains several million status updates for the years 2020 and 2021 and should give a credible insight into user emotions during Covid19.

Table 4. Distribution of tweets in the dataset by year.

First Evaluation					Current Evaluation	
2015	2016	2017	2018	2019	2020	2021
138.984	134.807	273.948	1.512.112	8.128.685	6.482.840	4.627.022

Analysing the average emotions of all users over time

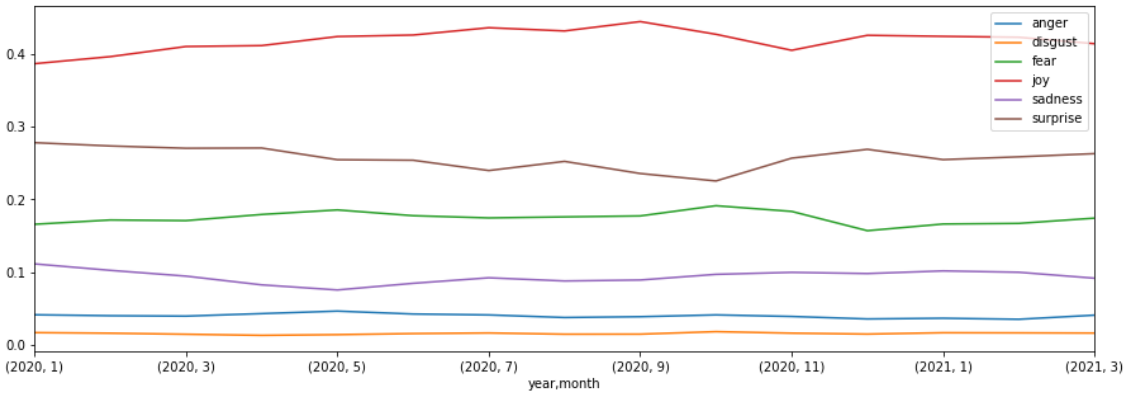
The first aggregation contains all user tweets and the associated emotions for each year. Table 5 shows the mean values for each emotion with joy and surprise as the dominant emotions, joy even more so for the years 2020 and 2021. It is noticeable that sadness is declining while anger and fear are rising until 2019 and then also declining.

Table 5. Average Emotions of all users per year.

Year	Anger (%)	Disgust (%)	Fear (%)	Joy (%)	Sadness (%)	Surprise (%)
2015	3,7	1,9	14,7	32,1	14,6	33,0
2016	4,5	1,9	16,4	32,8	12,6	31,8
2017	5,1	1,6	16,5	34,9	11,7	30,2
2018	5,4	1,8	17,6	35,1	11,4	28,7
2019	5,8	2,0	18,0	32,9	11,8	29,6
2020	4,0	1,6	17,6	42,2	9,3	25,3
2021	3,7	1,7	16,9	42,0	9,8	25,8

Figure 2 shows the user emotions for each analyzed month. It can be seen that fear rises when the first cases of Covid19 turn into a pandemic around March 2020 and before the United States presidential election in November 2020. Notably, expressed joy is rising throughout 2020 and just drops when the presidential election concludes, while surprise increases afterwards.

Figure 2. User emotions grouped by month.



Analyzing the emotions of user groups based on the user´s followers

The distribution of twitter users below is based on the number of followers a user has. The mean and median differ, which stems from outliers with an extreme amount of followers. The following analysis categorizes users based on the quartiles of this distribution. The amount of followers an average user has increased between the first analysis and the current analysis for all quartiles.

Table 6. Distribution based on the amount of followers.

Year	Mean	Std	Min	25%	50%	75%	Max
2015-2019	3.466	40.188	0	67	386	1.407	1.653.827
2020-2021	5.069	38.146	0	224	765	2.255	1.754.784

The distribution of emotions with user groups based on the amount of followers the users have, shows clear distinctions between the groups for all emotions in 2015-2019 and for all groups but anger for 2020-2021, e.g., expressed fear and joy continually decreases from the users with the least amount of followers to the users with the most followers. The emotions disgust, sadness and surprise are expressed more strongly the more followers a user has, with the highest values for the user group with >75% of followers.

Table 7. Users with less than 25% of followers.

2015-2019 / 2020-2021	Anger (%)		Disgust (%)		Fear (%)		Joy (%)		Sadness (%)		Surprise (%)	
Mean	7,5	3,8	1,0	1,0	20,5	18,0	36,2	48,0	7,7	4,3	27,0	24,9
Std	13,4	6,5	2,4	3,6	17,9	19,6	24,4	31,1	11,6	9,3	21,7	27,3
Min	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
25%	1,3	0,2	0,2	0,0	7,0	4,0	16,4	19,9	1,6	0,1	11,3	1,9
50%	2,9	1,7	0,5	0,1	15,4	11,2	32,7	45,4	4,1	0,9	21,0	15,5
75%	7,1	4,9	1,1	0,5	29,1	24,8	51,6	75,0	9,1	3,7	37,6	38,9
Max	100	100	96,9	98,5	100	100	100	100	100	100	100	100

Table 8. Users with followers >25% and <50%

2015-2019 / 2020-2021	Anger (%)		Disgust (%)		Fear (%)		Joy (%)		Sadness (%)		Surprise (%)	
Mean	5,8	3,9	1,9	1,3	17,7	17,8	33,9	45,4	11,5	6,5	29,0	25,1
Std	11,0	7,0	4,2	4,2	18,2	19,9	27,8	31,8	16,3	12,6	24,2	26,8
Min	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
25%	0,8	0,3	0,2	0,0	4,6	3,6	11,1	16,1	1,8	0,2	9,2	2,4
50%	2,4	1,7	0,7	0,1	11,8	10,8	26,1	41,2	6,0	1,3	23,3	16,3
75%	5,5	4,6	1,9	0,9	24,4	24,5	50,6	72,7	14,1	7,1	44,4	39,4
Max	100	100	99,5	98,4	100	100	100	100	100	100	100	100

Table 9. Users with followers >50% and <75%.

2015-2019 / 2020-2021	Anger (%)		Disgust (%)		Fear (%)		Joy (%)		Sadness (%)		Surprise (%)	
Mean	5,4	4,0	2,1	1,7	17,2	17,3	33,0	40,6	12,4	10,5	29,8	25,9
Std	10,3	8,0	4,6	4,8	18,3	20,2	28,2	32,2	17,1	16,0	24,8	25,7
Min	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
25%	0,7	0,3	0,2	0,0	4,4	3,2	10,1	11,6	1,9	0,5	9,1	3,7
50%	2,2	1,6	0,8	0,4	11,0	10,0	24,2	32,5	6,6	4,1	24,6	18,5
75%	5,3	4,2	2,1	1,5	23,4	23,6	49,2	66,9	15,1	13,8	46,3	40,6
Max	100	100	99,5	98,7	100	100	100	100	100	100	100	100

Table 10. Users with followers >75%.

2015-2019 / 2020-2021	Anger (%)		Disgust (%)		Fear (%)		Joy (%)		Sadness (%)		Surprise (%)	
Mean	5,0	3,9	2,3	1,9	16,9	16,9	31,1	40,2	13,7	11,5	31,1	25,7
Std	9,2	8,2	4,6	5,0	17,7	20,2	26,6	32,6	17,3	16,5	24,2	25,1
Min	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
25%	0,8	0,3	0,3	0,1	4,6	2,9	9,8	11,2	2,8	0,9	11,5	3,9
50%	2,3	1,4	0,9	0,5	11,0	9,4	22,8	30,9	8,0	5,3	26,4	18,8
75%	5,0	3,8	2,4	1,7	22,8	22,8	46,0	67,2	16,9	14,9	47,2	40,5
Max	100,0	100	99,3	99,1	100	100	100	100	100	100	100	100

Average emotions of user groups per year

The following tables show the analysis of all tweets of a user group and the resulting mean values per emotion for a certain year. Joy and surprise are the emotions with the highest mean across all user groups and years. The users with the least amount of followers (<25%) expressed anger, fear and joy most strongly in the years 2015-2019. Joy was communicated even more strongly across all user groups in 2020 and 2021. Contrary to this, the expressed sadness and disgust increases the more followers a user has. It is also noticeable that across all user groups the values for surprise and sadness decrease over the years.

Table 11. Less than 25% of followers.

Year	Anger (%)	Disgust (%)	Fear (%)	Joy (%)	Sadness (%)	Surprise (%)
2015	6,6	0,9	19,6	36,7	8,6	27,6
2016	6,7	0,9	18,6	34,3	7,5	32,0
2017	6,2	1,0	18,6	37,4	8,5	28,3
2018	6,4	1,0	19,4	37,5	7,9	27,9
2019	7,9	1,0	20,9	35,9	7,6	26,7
2020	4,0	0,9	18,4	47,9	4,3	24,4
2021	3,4	1,0	16,8	48,2	4,3	26,4

Table 12. Followers >25% and <50%.

Year	Anger (%)	Disgust (%)	Fear (%)	Joy (%)	Sadness (%)	Surprise (%)
2015	3,6	1,7	14,2	32,6	14,2	33,8
2016	4,9	1,8	17,5	33,4	11,6	30,7
2017	5,1	1,6	16,4	35,2	11,6	30,1
2018	5,3	1,9	17,6	35,2	11,5	28,5
2019	5,9	1,9	17,8	33,7	11,5	29,1
2020	4,0	1,3	18,3	45,1	6,7	24,6
2021	3,6	1,1	16,8	46,1	6,1	26,3

Table 13. Followers >50% and <75%.

Year	Anger (%)	Disgust (%)	Fear (%)	Joy (%)	Sadness (%)	Surprise (%)
2015	3,2	1,8	13,7	32,8	15,4	33,1
2016	5,0	1,8	18,1	34,0	11,5	29,5
2017	4,9	1,7	15,5	35,1	12,2	30,6
2018	5,2	2,0	17,2	34,5	12,1	28,9
2019	5,5	2,1	17,3	32,7	12,4	30,0
2020	4,1	1,7	17,5	39,9	10,8	26,0
2021	3,8	1,7	17,1	41,9	10,0	25,5

Table 14. Followers >75%.

Year	Anger(%)	Disgust (%)	Fear (%)	Joy (%)	Sadness (%)	Surprise (%)
2015	3,5	2,1	14,6	31,5	15,3	32,9
2016	3,9	2,1	15,5	32,2	13,9	32,4
2017	4,4	2,1	15,3	32,7	13,9	31,6
2018	5,1	2,1	16,4	33,6	13,3	29,5
2019	5,5	2,3	17,1	30,6	13,7	31,2
2020	4,0	1,8	17,0	40,2	11,5	25,5
2021	3,8	1,9	16,8	40,2	11,5	25,8

DISCUSSION

The implemented real-time data analysis pipeline collected 11 million tweets from roughly 6000 users over the time span of several weeks. The analysis of this dataset highlights joy and surprise as the most expressed emotions among all users and all years, with joy even increasing in 2020 and 2021. The evaluation of the collected data without user groups shows that expressed sadness and surprise decline each year among all users. It is reasonable to think that the election has a major impact on the outcome, as suggested in the aggregated monthly data.

The users have been categorized for a follow-up analysis based on their interaction with the social media platform and each other to get more detailed insights. The amount of followers a given user has was used as metric, that describes a user's reach on the platform and the interest other users have in her or him. Classifying the users based on the amount of followers showed differences between expressed emotions of user groups and suggests that this criterion is characteristic. The results allow the evaluation of the constructed hypothesis as follows:

- The expressed joy (H1) is strongly increased for 2020 and 2021 in comparison to the earlier timespan from 2015-2019. This is true for the aggregated mean of all users but also for each of the individual user groups. Thus, H1 can be rejected.

- The negative emotions, i.e., anger, disgust, fear and sadness, but also surprise, decline in 2020/2021 despite Covid19. Thus, H2-H6 can be refuted.
- The analysis of user groups based on the amount of followers clearly differentiates the groups, just as in previous work. Despite Covid19 the values for expressed emotions clearly increase / decrease from each group to the next and it is possible to verify H7.

CONCLUSION

The analysis of ~11 million tweets by 6000 users over a time period of a few weeks allows several conclusions. The KOARCH architecture enabled us to easily setup a big data pipeline, consisting of virtualized building blocks, with continuous operation under high load.

For the analysis users have been grouped based on the number of followers they have.

The prevalent expressed emotions on twitter were joy and surprise. Over the observed period of time, from 2015 to 2019, the values for joy remained consistent, while an increase in anger, disgust and fear could be verified for all user groups. Sadness on the other hand declined, maybe it was transformed into anger or fear. It is noticeable that twitter users with the least amount of followers (<25%) expressed anger and fear most strongly. However, for 2020 and 2021 during the Covid19 pandemic and the United States election the joy was massively increased while the negative emotions declined. Aggregating the data for the individual months helped to show differences that could be explained with the election.

There are several limitations to this study. Access through the official twitter interface was limited to the last 3.200 tweets of any given user. While this is enough for the majority of users, for some of the power users this was just a fraction of their status update history. Collecting data several times, as done for this study, and combining the resulting datasets avoids this limitation. Predictions derived from a machine learning model are always just as good as the underlying model. Even though the results of Colneric and Demsar are really impressive, there may be a bias towards a certain emotion in their model. It is also notable, that this model was trained on a massive dataset of English text, but therefore it can be used to classify text written in English only.

It would be interesting to conduct the same analysis for the rest of 2021 and the following years to see if the recovery from Covid19 or the political scenario further influences the expressed emotions.

KEYWORDS: Social Media, Machine Learning, Natural Language Processing, Emotion Classification.

REFERENCES

Bamman, D., & Smith, N. A. (2015). Contextualized sarcasm detection on twitter. *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, 574-577.

Colneric, N., & Demsar, J. (2018). Emotion Recognition on Twitter: Comparative Study and Training a Unison Model. *IEEE Transactions on Affective Computing*, *3045*(c). https://doi.org/10.1109/TAFFC.2018.2807817

Ekman, P. (1999). Basic emotions. In *Handbook of Cognition and Emotion*. John Wiley & Sons Ltd.

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter Sentiment Classification using Distant Supervision*. 1-6. https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). *Identifying Sarcasm in Twitter: A Closer Look*. http://www.vidarholen.net/contents/interjections/

Hogue, J. V., & Mills, J. S. (2019). The effects of active social media engagement with peers on body image in young women. *Body Image*, *28*, 1-5. https://doi.org/10.1016/j.bodyim.2018.11.002

Hunt, M. G., Marx, R., Lipson, C., & Young, J. (2018). No more FOMO: Limiting social media decreases loneliness and depression. *Journal of Social and Clinical Psychology*, *37*(10), 751-768. https://doi.org/10.1521/jscp.2018.37.10.751

Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! *International AAAI Conference on Web and Social Media*.

Levenson, J. C., Shensa, A., Sidani, J. E., Colditz, J. B., & Primack, B. A. (2016). The association between social media use and sleep disturbance among young adults. *Preventive Medicine*, *85*, 36-41. https://doi.org/10.1016/j.ypmed.2016.01.001

McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). EITS Manual for the Profile of Mood States. *Educational and Industrial Testing Service*, *3*(27), 1984.

Mohammad, S. M., & Kiritchenko, S. (2015). Using Hashtags to Capture Fine Emotion Categories from Tweets. *Computational Intelligence*, *31*(2), 301-326. https://doi.org/10.1111/coin.12024

Nodarakis, N., Sioutas, S., Tsakalidis, A., & Tzimas, G. (2016). *Using Hadoop for Large Scale Analysis on Twitter: A Technical Report*. http://arxiv.org/abs/1602.01248

Plank, B., & Hovy, D. (2015). *Personality Traits on Twitter—or—How to Get 1,500 Personality Tests in a Week*. 92-98. https://doi.org/10.18653/V1/W15-2913

Plutchik, R. (1982). A psychoevolutionary theory of emotions. *Social Science Information*, *21*(4), 529-553.

Strohschein, J., Lara Palma, A. M., & Faeskorn-Woyke, H. (2019). Detecting emotions in social media. a technological challenge to enhance youngest behavior. *28th AEDEM International Conference - Management in a Smart Society: Business and Technological Challenges*.

Tasoulis, S. K., Vrahatis, A. G., Georgakopoulos, S. V., & Plagianakos, V. P. (2018). Real Time Sentiment Change Detection of Twitter Data Streams. *2018 IEEE (SMC) International Conference on Innovations in Intelligent Systems and Applications, INISTA 2018*, *Us*. https://doi.org/10.1109/INISTA.2018.8466326

HOW A BRAIN-MACHINE INTERFACE CAN BE HELPFUL FOR PEOPLE WITH DISABILITIES? VIEWS FROM SOCIAL WELFARE PROFESSIONALS

Yohko Orito, Tomonori Yamamoto, Hidenobu Sai, Kiyoshi Murata, Yasunori Fukuta, Taichi Isobe, Masashi Hori

Ehime University (Japan), Ehime University (Japan), Ehime University (Japan), Meiji University (Japan), Meiji University (Japan), Health Sciences University of Hokkaido (Japan), Waseda University (Japan)

orito.yohko.mm@ehime-u.ac.jp; yamamoto.tomonori.mh@ehime-u.ac.jp; sai.hidenobu.mk@ehime-u.ac.jp; kmurata@meiji.ac.jp;yasufkt@meiji.ac.jp; tisobe@hoku-iryo-u.ac.jp; horimasa@waseda.jp

ABSTRACT

A brain-machine interface (BMI), one of a group of emerging cyborg devices, processes signals acquired from a human brain and translates them into a meaningful output in accordance with a given purpose such as operating a machine remotely. Thanks to these functions, BMI systems are expected to be adopted to support people with disabilities, in particular for those who cannot move their body by their own intentions. However, the ethical issues and social concerns surrounding the use of BMI systems for people with disabilities have not been exhaustively discussed. The social and ethical implications for social welfare should be carefully considered in advance of the actual use of BMI, to secure users' well-being. In this study, the authors conducted experiments on two healthy social welfare professionals, with specialised knowledge and experience, as subjects using a non-invasive wearable BMI device. The subjects completed an interview survey, before, during, after the experiment, to investigate the availability, usability and ethical issues of BMI systems for people with disabilities.

INTRODUCTION

Cyborg technologies are regarded as useful tools that can promote daily quality of life and enhance physical and intellectual abilities. Currently, such technologies are extremely useful in assisting people with disabilities. Among them, the technological development of a brain-machine interface (BMI), or a brain-computer interface (BCI), has shown remarkable progress over recent years, and its utilisation by people with disabilities has steadily been trialled and adopted in developed countries. A BMI promotes communication through signalling from the brain to external devices and vice versa using dedicated hardware and software (Orito et al., 2020). Using these functions, BMI has been adopted to support people with disabilities, for example, a BMI device enables individuals who are unable to move their bodies at will to successfully control digital devices and to communicate with others.

However, discussions regarding the utility and ethical issues surrounding the use of BMI systems in social welfare have not been extensive; this is because BMI devices or systems are not currently widely used in the daily lives of people with disabilities. Not only the physical and psychological risks, but also the social issues and ethical implications for social welfare should carefully be considered in advance of the actual use of BMI, in order to secure users' well-being. Furthermore, it is a common tendency for healthy people to consider the simple-minded arguments or benefits of cyborg technologies that can be used for supporting people with disabilities and to miss the disadvantages and risks that may

take place through the use of such technologies. Of course, there are many beneficial ways in which BMI systems can assist people with disabilities; however, we should be mindful not to overestimate these beneficial aspects. To this end, BMI usability for people with disabilities, the appropriateness of targeted subjects, its purpose, and the risks and ethical issues entailed in the operation of these systems should be proactively examined from various viewpoints.

With this in mind, the authors conducted experiments on two healthy social welfare professionals as subjects through fitting them with a non-invasive wearable BMI device and used an interview survey to investigate the availability, usability and ethical issues of BMI systems for people with disabilities. This research is in the exploratory phase; as such, it is meaningful to ask the views and opinions of professionals or experts in related areas. In these experiments, the subjects were asked to put the non-invasive BMI device (Electro Encephalo Graphy input device) on their head and to operate a robotic arm remotely or without touching it, as if using psychokinesis, as in our previous study (Orito et al., 2020; Orito, Murata and Suzuki, 2020). Moreover, based on the results of these and other studies by the authors (Isobe, 2013; Murata et al., 2017, 2018, 2019), we conducted interviews with the two subjects, before, during and after experiments, to examine their attitudes to BMI device usage, feelings regarding robotic arm operation, ethical awareness of brain signal collection by a BMI device, and to estimate the benefits and risks surrounding the utilisation of BMI systems for people with disabilities.

OVERVIEW OF THE BMI EXPERIMENT AND THE INTERVIEW SURVEY

Outline of the experiment and the interview

The experiment was designed so that the subjects controlled a robotic arm using only their brain signals via a non-invasive wearable BMI/EEG (Electro Encephalo Graphy) device placed on their head. During the training stage of the experiment, the subjects' brain signals were collected and recorded by the system through the BMI/EEG devices, then two types of brain signal data—a 'relaxed state' and an 'in-operation state'—were recorded by the systems, and the data were transmitted to and stored in the application software.

After the training, each subject was asked to intentionally move to 'in-operation state' by imaging that they were pushing the robotic arm backward. If their brain signal patterns were matched with the in-operation state registered in the software system, the robotic arm turned toward the back (for detail information regarding the experimental procedure, see Orito et al., 2020; Orito, Murata and Suzuki, 2020).

Survey participants

The experiments and interviews were conducted in August 2020 at Ehime University in Matsuyama, Japan. All procedures performed for this study were in accordance with the ethical standards of the research ethics committee at the Faculty of Medicine, Ehime University. The attributes of two subjects of the experiment are shown in Table 1. They were healthy adult professionals; one was a researcher majoring in social security law and social welfare, and the other was a social worker. The subjects were knowledgeable regarding BMI technologies, on the basis of their professional experiences.

Each subject was informed of the purposes and methodology of this study and the contact information for enquiry in advance of the experiment. Based on this informed consent process and agreement, before, during and after the experiment, the subjects were asked to grant semi-structured interviews, or to respond to interview questions in written form at a later date. These interview questions were designed so as to examine subjects' attitudes to and recognitions of their experience with the BMI

system during the experiments, based on previous studies (see, Orito et al., 2020), and these questions were categorised as follows: (a) privacy and personal data protection, (b) human autonomy and dignity, (c) identity development and personal transformation, (d) the acceptance of body extension in an individual and organisational context, (e) workplace cyborgisation and (f) social responsibility and informed consent. In addition to these, further questions related to the BMI experiment and support for people with disabilities were also included. Throughout the experiments, subsequent interviews and follow-up surveys (via by e-mail after the experiment) regarding BMI devices and their usage, the authors gained insightful findings as explained in the next section.

Table 1. Subjects of the experiment (n = 2).

ID	Age	Gender	Occupation	Have you ever heard about a BMI or are you familiar with a BMI?	Expectation/anxiety about the experiment (Weak 0–Strong 7)
1	40s	Female	Professor (major in social security law)	Yes. While I didn't know the name of the system, I had heard that ALS (Amyotrophic Lateral Sclerosis) and muscular dystrophy patients used such systems. These systems had attracted attention at international exhibitions concerning medical and welfare apparatus, several years ago.	3/4
2	40s	Male	Social worker	Yes. I heard about this type of technology in our communities, several years ago. However, at that time, it seemed that such systems were unsatisfactory in supporting decision making for people with disabilities, due to technological difficulties.	7/1

RESULTS OF THE INTERVIEW SURVEY

Beneficiaries of the BMI systems and suitable subjects of the experiment

The two interviewees (IDs 1 and 2) were asked to provide us with their opinions on what kind of people with disabilities would receive the benefits from the usage of BMI systems, and the medical and other conditions to be satisfied to let them participate in this BMI-system-based experiment as subjects. The interviewees suggested that the BMI systems and the experiments were available or useful for people with ALS, muscular dystrophy, cervical spine injury, cerebral palsy, myasthenia gravida, and so on. Furthermore, it was suggested that people who had acquired disabilities would be more suitable as subjects rather than individuals with congenital disabilities. These points were explained as follows.

> '*It is possible that individuals with muscular dystrophy or ALS (Amyotrophic Lateral Sclerosis) could use this system. In the case of ALS, many of these patients become disabled since they were grown, so I think that they might be suitable for the use of this system, depending on their circumstances. People with muscular dystrophy and ALS are comparatively less likely to develop disabilities concerning their ears or intellects although this does not mean all patients with muscular dystrophy and ALS will not develop problems in their participation of the experiment. The one problem that people with ALS and muscular dystrophy would have with this experiment is in articulating their response; otherwise, there should be no other challenges. Apart from*

taking time to explain the details of the experiment or research, there should be no need for researchers to provide any additional training for these people in order to accommodate them as participants in an experiment'. (ID 1)

'*If it is possible to indicate their intentions using a robotic arm according to their desire, this may be a good system that can be usable for not only people with ALS but those who have had cervical spine injury. In cervical spine injury cases, if the nerve in the neck is damaged, the lower part of the neck becomes motionless from that point downwards; in addition, sometimes their hand movements and grip strength decrease significantly. Therefore, if the robotic arm can grab something according to their own will without using their hand or other body parts, some such people will view this positively and it will give them hope. It is also applicable for people with cerebral palsy and those with paediatric palsy*'. (ID 2)

'*Myasthenia gravis is an intractable disease that gradually makes it impossible for patients with it to move the muscles of their whole body. When such individuals are keeping good conditions, they can move independently, but during a strong malaise they easily become tired, and it becomes exhausting to move by themselves. Therefore, if a robotic device could be operated with one's head and brain signals, this would open up considerable prospects for such people*'. (ID 2)

Moreover, the social worker (ID 2) pointed out several notanda when the experiment is conducted with people with disabilities as subjects.

'*Wouldn't it be better for people with disabilities to work with the imagery of 'moving the objects' rather than 'pushing the objects'? Perhaps, receiving feedback from the people with disabilities after having them try out the experiment might give clarification on this point*'. (ID 2)

'(After the experiment) *When people with disabilities participate in this experiment, in some cases it may be difficult for them to wear the headset, depending on the angle of the wheelchair and other support devices. Some people are in a constant state of apparent sleep and some cannot make their bodies wake up. It is common for individuals with cerebral palsy to have involuntary movements which may dislodge a headset; it would therefore need to be tightly fitted so that it does not come off their heads*'. (ID 2)

Possibilities and value of BMI usage for people with disabilities

Many ideas regarding the possibilities and future values of BMI usage for people with disabilities were described by the two interviewees. In particular, both of them suggested its usefulness for assisting or supporting the physical movements of people with disabilities. The importance of daily assistance for them to maintain their human dignity and a sense of their self-reliance was also suggested by the interviewees.

'*I personally believe that there is no question that those with disabilities would prefer to be able to go to the bathroom on their own. This ability is related to their dignity as a human being. For those who cannot do this on their own, I think there would be a high demand for a robotic arm, if one can use it to do things like going to the toilet, pulling down one's pants, and wiping one's bottom. Such uses for a robotic arm are important for people with disabilities; this may be a more fundamental need than their social participation. It is said that putting on a diaper is one of the most humiliatingly difficult things for individuals with a disability as they get older; thus, I believe that people would prefer to be able to go to the bathroom on their own*'. (ID 1)

'*Rather than using BMI for working or getting a job, I guess, the needs for people with disabilities to use such a device are more closely related to their ADL (Activity of Daily Life), such as using the toilet for themselves or masturbating. Regarding masturbation, it is rare that a person with disabilities would receive help with that, at least in Japan; and people whose hands are paralysed cannot do it. Therefore, I think there is a need for automated assistance under such personal situations*'. (ID 1)

'*I think that everyone has a strong desire for self-reliance of being able to do everything in daily life by oneself, although no one lives completely alone. When I asked people with disabilities about the accessibility of public transportation, a typical response was "my goal is to be able to use public transportation as easily as a healthy person, without special help from anyone else". If the goal is to reduce the physical care of others, even a little, I believe that technology such as a robotic arm is highly desirable for people with physical disabilities because their actions can be performed by themselves, without any assistance by others, by using such a machine. Sometimes I tell people with physical, intellectual or mental disabilities, during my work, that being able to rely on others is "an excellent ability" and "a life is designed so that it should not be lived alone". Even if everyone knows this, it is still difficult for them to actually understand it in their minds. However, having been in communication with people with disabilities for a long time, I have found that there are individuals who find their own way to reconcile society as well as others. The encounter with those people is truly a precious moment for me. However, I think it's very hard for us to completely free ourselves from meritocratic thoughts*'. (ID 2)

Interviewee 2 also emphasised that these technologies could be useful to display an indication of their intention for people who cannot move their hands or eyes by themselves.

'*Regarding moving a machine with brain signals, for those individuals who will totally lose the capacity to move their body, I think that it is important for them to train in the use of the machine on a regular basis while they still have some control of their movement. For example, when we talk with people who express their intentions with eyelids only, sometime their intentions can be misunderstood. It would be better if there was something that could somehow confirm their true intention. For people who are locked-in or cannot move any part of their body, a device that enables them to do something with their brain signals is the only hope. The successful adoption of such a device for their daily life sounds great. The quality of the usage of such a device could be improved through a trial and error approach*'. (ID 2)

Furthermore, he explained the support required by people with severe mental disabilities and suggested the usefulness of BMI devices for them as follows.

> '*When we support a child with severe mental disabilities, who cannot talk with others, and something goes wrong with them, we – care workers – guess what this person would talk if he/she could, and decide to do something for him/her under an agreement to a certain extent between us. If we could know clearly that he/she wants to say through the use of this device, it would be really helpful for him/her as well as for us*'. (ID 2)

> '*I suppose that the use of the brain signals of those patients who have severe motor and intellectual disabilities in their daily lives is helpful for them. Such patients tend to be considered cerebrally handicapped to the eye, due to their inaction. But, some of them may be mentally healthy and sensible even though they cannot, for example, consciously move their eyes. If such a physical and mental state can be detected through measuring brain signals, this is really nice. Actually, hearing impairment is sometimes judged by the response of brain signals to sound*'. (ID 2)

On the other hand, Interviewee 1 suggested that research into cyborg technologies, in particular BMI, could possibly be used during palliative care. She explained as follows.

> '*I actually knew a person who had phantom pain, the pain was excruciating the person suffering from it. Once, I took time to consider what happens to brain signals of the person when this pain occurs. I understand that nothing can be done for someone suffering such pain in a part of the body that is not physically present, but to the person, the feeling of pain is real; thus, it may be an imaginary pain. Based on the research regarding cyborg or BMI technologies, if such a mental mechanism can be analysed, it would be useful for the advancement of palliative care. I think, this is a desirable technology application, which could respond to the particular need of pain relief using brain functions*'. (ID 1)

The importance of the autonomy of people with disabilities in the use of cyborg technologies

Both of the interviewees emphasised the importance of an independent decision made by individuals with disabilities as to the use of supportive cyborg technologies. In response to the question 'What conditions should be satisfied for the use of BMI by those with disabilities to be socially acceptable?' Interviewee 1 pointed out that '*The desire and choices of the person concerned are matters, rather than the degree of his/her disabilities*', and showed the conditions that were needed in order for people with disabilities to take advantage of cyborg technology such as BMI systems as follows.

> '*Around 2005, when I went to Denmark, I saw that people with disabilities were able to use tools that enabled them to engage in a more diverse range of communication than they would have been able to had they been in Japan. At that time, it was common to see people using mobile phones and machinery to express their intentions. Therefore, I understood that if the coordinators and the people themselves so wished, and if they had the opportunity, they would*

be able to communicate better. We discussed about how it was important to have a diverse range of options'. (ID 1)

'*For people with progressive intractable diseases and physical disabilities, to improve their ADL, the use this kind of technology should be conditioned on that those people express their wish to use it and a person in charge of the management of care of them recommends; I think those people can use the BMI machine. On the other hand, however, it may be difficult to confirm their intentions, depending on the situation*'. (ID 1)

Interviewee 2 also recognised the importance of confirming the intentions of people with disabilities to use cyborg technologies.

'*If cyborg technologies are used as a tool for decision-making, it will be necessary to establish a manual or an ethical procedure to confirm users' will. Even if a device which allows person with severe mental and physical disabilities to express their intentions is developed, regulation or a mechanism to ensure the expressed intentions are true has to be established*'. (ID 2)

Risks and ethical concerns regarding the use of cyborg technologies and BMI systems for people with disabilities

Malfunctions and over-adaptations of BMI systems

Needless to say, as long as BMI systems are technology-based ones, they are not completely free from malfunction. The fear of this was explained by the two interviewees as follows.

'*For anyone, including me and people with disabilities, BMI system usage inevitably entail risks such as systems' becoming out of control and a third party's control over users*'. (ID 1)

(Responding to a follow-up survey question via email) '*The human will is vague, and hard to understand. One's words and deeds are not necessarily identical with one's intentions. But, if it is almost totally proved by measuring one's brain signals that one's brain surely ordered a certain deed which one considers the result of the malfunction of BMI systems, that deed is no longer the outcome of the system malfunction, but was based on one's true intentions. Even when making a simple either-or choice, a similar situation would happen. Then, we need to clarify what malfunction is. It may be necessary for us to set a way to measure the degree of malfunction. Of course, I think it's very hard to investigate every element that caused malfunction, because restaging the situation where the malfunction happened is totally difficult*'. (ID 2)

(About implantable brain chips) '*Recently, implantable electrodes have been used to control epileptic seizures for patients with cerebral palsy, in particular. However, this treatment sometimes cause fear. In the field of psychiatry, encephalotomy is applied to the treatment of*

epilepsy, but this would be associated with electric shock treatment or something similar. Given these, implantable BMI would arouse a feeling of fear because electrodes are implanted in the brain. For me, implanting something in the brain itself is scary'. (ID 2）

Moreover, Interviewee 2 had concern about the negative effect on mental status of system users whose brain functions might be taken over by the machines.

'*A challenge is how voluntarily one can choose the ways of processing the information acquired through the use of implantable BMI, regardless of whether one has a disability or not. If such a device allow people to experience hallucination or auditory one, which is a psychiatric symptom, users of the device may not be able to differentiate hallucination from a reality. Given this, I fear about what consequence would be led to by the use of implantable BMI early in people's life*'. (ID 2)

'*Among those who suffer mental disorder, there are some people who experience thought withdrawal, thought broadcasting or another type of delusion – for example, they believe they are under 24/7 monitoring and their brain information are collected by police. These symptoms disturb their daily life. I have usually told such people tenderly that no one can do such things, or doing those things is prohibitively costly even if technologically possible, and the police is not interested in your private life. However, the development of a technology which enable anyone to do those things would result in worsening of symptoms of such patients. In the field of psychiatry, brain surgery to implant a BMI device would be associated with lobotomy or electric shock therapy, which would call up a bad image. Psychiatric medicine is adjusting the balance of substances in the brain, so it touches the brain in a sense. Even now, some people complain that they are controlled by me (through medicine)*'. (ID 2)

'*As with alcohol, drugs and games, each of which has an addictive nature, once people started to use (cyborg) equipment daily, they would get addicted to it more or less. I'm not sure whether they would consider that it's okay for them not to use the convenient equipment or morbidly feel restless, if such daily use is terminated*'. (ID 2)

The transformation of social recognition of people with disabilities

Both of interviewees indicated the possibility that public attitudes toward people with disabilities and the concept of assistance to them could be transformed by the implantation of cyborg technology. They recognise that these are closely associated with how the public view, accept, and respect people with disabilities and their dignity. The following quotes aptly illustrate this point:

'*From the 1990s, in the field of social welfare, latest technologies have been used with the consent of those with disabilities to maintain and promote their right to self-determination. However, it has been argued that the development of technologies focused on reducing the burden of care-givers is problematic because it may lead to the use of the technologies to monitor care receivers and the invasion of personal rights of them. It is also feared that technology may be introduced in places where the will of individuals with disabilities is not be*

considered. Another inhumane way of using such technologies is that if those with disabilities didn't wish to use the technologies, then any care would not be provided to them. The latter topic is more seriously debated than the former'. (ID 1)

'*It may be possible to overcome the difficulties those who have disabilities have in daily life through implanting cyborg devices. However, considering that disability has only recently been re-regarded as a social issue rather than an individual issue, I worry that the use of such cyborg devices may lead to social tendencies to re-perceive disability as a problem of the individual*'. (ID 2)

'It is important for us NOT to create a situation, where those who have disabilities and use cyborg devices are treated as machines or not-genuine humans. They should be treated as human-beings. It is important to create a culture where the usage of cyborg devices by individuals is considered completely natural; the environment in which it is widely accepted that inconspicuous cyborg devices are indispensable in maintaining users' human dignity while the prejudice and discrimination against those users are eliminated should be developed. Those who aspire to use cyborg devices should be able to use such devices out of their own will. They should free themselves from any stigmas associated with cyborg technologies and be ready to lead a life as human beings confidently'. (ID 2)

Equal opportunity

Discussions concerning the use of cyborg technology, such as BMI systems, involve other ethical and social issues. Equal opportunity in becoming cyborg is a typical one. The widespread adoption of cyborg technology would result in widening the gap between the haves and have-nots. Should the equal opportunity of becoming a cyborg be ensured for all citizens?

'*It is conceivable that a new socioeconomic class would be created among non-disabled people, for example the blue-collar cyborg class and the white-collar non-cyborg class. This would lead to the situation where job opportunities across classes are extremely limited, and economic disparities between classes are entrenched. Such disparities would be created, in the same way, between those with and without disabilities. However, I'm afraid that policies in terms of the development and implementation of new technologies including cyborg technologies would not be designed so as to ensure equality among all citizens*'. (ID 1)

'*In general, welfare equipment is very expensive, especially in its introductory period. Thus, unless the government provides financial support to those who necessary to use such equipment, the use of it would be limited to those who afford to buy it. So, only wealthy people would be able to use cyborg devices. Poor people with disabilities would have no chance to access the devices even though they have the same disabilities as the rich*'. (ID 2).

The similar situation would arise in workplace. Interviewee 2 mentioned this as follows.

'*In workplace, what if those employees who implant cyborg devices in their bodies receive a higher job performance evaluation, thanks to the devices, than those who don't? This is possible, given the fact that cyborg devices enhance people's abilities. Then, employers may recommend their employees to implant cyborg devices. Similar to the case of people with disabilities, the apparent ability gap owing to cyborg devices between cyborgs and non-cyborgs may let employees have an obsessive-compulsive desire to implant cyborg devices in their bodies. They effectively have only one choice if they want to feel themselves "normal" in their workplace. I would like to work in such workplace, only if all of employees respect each person's freedom to become a cyborg*'. (ID 2)

Interviewee 1 also suggested the issue on the working condition, '*For those who are engaged in intellectual work using cyborg devices, their work time will be extended (because the devices enable to remove their feeling of fatigue). I feel uncomfortable about this*'. (ID 1).

Privacy and personal data protection

Interviewee 1 showed a strong fear about the risks caused by the use of BMI systems related to privacy and personal data protection.

'*I don't understand what signals are sent from my brain. On the other hand, given that the brain is an important organ for human intelligence and reasoning, I instinctively feel a fear, a modernistic fear, about being collected my brain signal data. If a person who measures my brain signals or explain the collection of my brain signal data is not trustworthy, I decline his/her getting my brain signals. Actually, I worry about monitoring people based on the results of analysis of brain signal data collected from those individuals who, for example, developed an addiction, committed a crime or reoffended after leaving prison. Imagine you are told "You have a brain signal pattern which demonstrates you are prone to commit a sexual offence". This is really scary. As with gene information, this type of information must be used in an ethical manner; brain signal information should never be abused*'. (ID 1)

(Responding to the question "Do you think that brain signal data the EEG collects need to be protected carefully as sensitive personal data? What do you think if your brain signal data are utilised for lie detection?") '*The risks related to these questions depend on how widely this kind of information can be used and in fact for what purpose the information is utilised. Even if a person thinks that this is quite sensitive personal information, policy makers or those who are interested in using it may tell that this information is just of brain signals and using this is not a serious issue. However, I feel it is this personal information that identifies self*'. (ID 1)

'*There is quite a big difference in people's attitudes to the risks, in particular, between those who imagine that this information can be used for purposes other than the original intent and those who don't. The quality and performance of cyborg devices are not a matter. Rather, the general-purpose properties of such information is a problem. It is scary if it is used politically, in particular for criminal policies. It would also be uncomfortable if such data were used in*

> *marketing to, for example, give customers recommendations based on their preferences specified by their brain signals*'. (ID 1)

On the other hand, Interviewee 2 presented a positive stance with the same questions on the protection of brain signal data as personal data.

> '*If the brain signals are used to identify individuals, this is interesting rather than scary. If the brain signals correctly indicate an individual's condition, psychiatrists feel that their work becomes easier*'. (ID 2)

DISCUSSIONS

The results of the experiments and subsequent interview surveys allowed the authors to recognise that subjects of the experiment should be selected based on symptoms people with disabilities have. In addition, it suggested that what purposes the BMI systems are effective for and what kinds of ethical concerns relating to BMI system use are important for them. The findings and insights gained from this experimental survey are summarised as follows.

- To conduct this type of experiment and survey in a continuous manner in order to analyse the utility of BMI systems and the ethical risks to people with disabilities, based on their appropriate informed consent for the survey, people with ALS, muscular dystrophy, cervical spine injury, cerebral palsy, myasthenia gravida and those who have acquired disabilities will be more appropriate as subjects.
- It seems that BMI systems or cyborg technologies may be useful for those who cannot move their body on their own to assist with their activities and communications in daily life, which may be important for respecting their human dignity and sense of self-reliance. Moreover, cyborg technology has the potential to assist people with severe mental disabilities in expressing their intentions or emotions; this also raises the possibility of its application during palliative care.
- Both of the interviewees emphasised that a critical factor is personal decision made by those with disabilities regarding whether they have the desire to use cyborg technologies such as BMI systems or not. In addition, it is necessary to demonstrate whether those people with disabilities are interested in using the cyborg technology or not, using proper measures through which they can clearly show their intention, for example, with their voice, gestures, or movement of their body parts in a certain manner. Even if they show the desire to use it, a preliminary training process is important for the usage.
- There are wide-ranging ethical and social issues surrounding the usage of cyborg technology, such as the BMI system, by people with disabilities. The issues are not limited to the risk of malfunction of those devices but include the mental health of their users, protection of user privacy and the potential changes it could bring to our society.
- While there are people who consider physical disabilities to be a personal problem rather than a societal challenge, the use of cyborg technology by those with disabilities could very well change our perception of the physically challenged population. Such a change in perception could stimulate discussion concerning diversity or ensuring equal opportunities for various

people. These issues must be carefully considered before the cyborg technology is adapted to social welfare services.

CONCLUSIONS

In this study, as a first step in examining the ethical issues of BMI usage for people with disabilities, experiments using BMI devices and interview surveys with two professionals from the social welfare sector were conducted. As in our previous study (Orito, et al., 2020; Orito, Murata and Suzuki, 2020), the number of subjects was small. However, the two interviewees provided useful and insightful suggestions for the development of more appropriate experimental environments, and the application of BMI systems or cyborg technologies in the social welfare sector.

On the other hand, when considering BMI systems for supporting people with disabilities, while the experimental survey and examination of professionals in social welfare was needed to analyse the practical and ethical issues in advance, the following question can be raised: 'Is it possible for a healthy person to understand and represent the feelings and sensations of people with disabilities, who cannot express their intentions or have limited physical sensation?' In other words, the experiences and results obtained from this experimental survey can evoke discussions based on the Eastern theory of body or Eastern theory of mind-body which deal with relationships between mind and body (e.g. Ichikawa, 1993; Yasuda, 2014; Yuasa, 1987; Yuasa, 1993). As a next step, along with examinations of the Eastern theory of body, implications regarding the use of cyborg technologies for people with disabilities as well as healthy people will be discussed.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers 20K01920 and 19K12528, the Kurata Grants subsidised by the Hitachi Global Foundation, and the Meiji University Grant-in-Aid for the international collaborative research project 'Cyborg Ethics'. We also appreciate Dr. Yoshitaka Moritsugu, and all the participants in the experiments and researchers who supported our study. We certify that all procedures performed in the experiments of this study, which involved human participants, were in accordance with the ethical standards of the research ethics committee established at the Faculty of Medicine, Ehime University (issued Jan.27.2020, No. 2001001).

KEYWORDS: brain-machine interface, support for people with disabilities, human dignity, privacy.

REFERENCES

Ichikawa, T. (1993). *The structure of "body" – Beyond the body theories* –. Tokyo: Kodansha (in Japanese).

Isobe, T. (2013). The perceptions of ELSI researchers to Brain-Machine Interface: Ethical & social issues and the relationship with society. *Journal of Information Studies*, 84, 47-63 (in Japanese).

Murata, K., Adams, A. A., Fukuta, Y., Orito, Y., Arias-Oliva, M. & Pelegrín-Borondo, J. (2017). From a science fiction to reality: Cyborg ethics in Japan. *Computers and Society*, 47(3), 72-85.

Murata, K., Fukuta, Y., Orito, Y., Adams, A. A., Arias-Oliva, M. & Pelegrín-Borondo, J. (2018). Cyborg athletes or technodoping: How far can people become cyborgs to play sports? Presented at ETHICOMP 2018, 25 September 2018, Retrieved from

https://www.researchgate.net/publication/327904976_Cyborg_Athletes_or_Technodoping_How_Far_Can_People_Become_Cyborgs_to_Play_Sports.

Murata, K., Arias-Oliva, M., & Pelegrín-Borondo, J. (2019). Cross-cultural study about cyborg market acceptance: Japan versus Spain. *European Research on Management and Business Economics*, 25(3), 129-137.

Orito, Y., Yamamoto, T., Sai, H., Murata, K., Fukuta, Y., Isobe, T. & Hori, M. (2020). The ethical aspects of a "psychokinesis machine": An experimental survey on the use of a brain-machine interface. In Arias-Oliva, M. et al. (Eds.), *Societal Challenges in the Smart Society Ethicomp book series* (pp. 81-91). Logroño, Spain: Universidad de La Rioja.

Orito, Y., Murata, K., & Suzuki, S. (2020). Possibilities and ethical issues surrounding brain-machine interfaces in the realm of social welfare: Potential for use by people with disabilities based on results from psychokinesis experiments. E-poster at the 68th academic conference of Japanese society for social welfare (in Japanese).

Yasuda, N. (2014). *The body of Japanese people*. Tokyo: Chikuma Shobo (in Japanese).

Yuasa, Y. (1987). *The body: Toward an Eastern mind-body theory*. Albany, NY: State University of New York Press.

Yuasa, Y. (1993). *The body, self-cultivation, and ki-energy*. Albany, NY: State University of New York Press.

MORAL DILEMMAS AND ETHICAL CONFLICTS RELATED TO MOBILE APPLICATIONS FOR SLEEP IMPROVEMENT

Ana María Lara-Palma, Montserrat Santamaría-Vázquez, Bruno Baruque-Zanón, Juan Hilario Ortiz-Huerta

Universidad de Burgos (Spain)

amlara@ubu.es; msvazquez@ubu.es; bbaruque@ubu.es; jhortiz@ubu.es

ABSTRACT

The sleep disorders are a variety of conditions that affect sleep and they seem to be increasing considerably among the population. At the same time, technological solutions have recently appeared claiming to help improve the sleep quality. Due to its ubiquity and ease of use, mobile applications are among the most commonly used technological aids to improve sleep among their users.

The objective of this work is to analyze, under an ethical judgement, the literature review about reliability and validity of mobile applications focused on sleep disorders, the ethical conflicts generated and the ethical judgment of the users.

The methodology followed consisted, on the one hand, an on-line search to collect a set of mobile applications focused on improving sleep that were recommended by some medical, governmental or at least non-commercial association or agency. The search has been conducted including the key words “sleep mobile app”. On the other hand, a search of scientific publications has been carried out in order to find objective evidence about the observable performance selected apps.

The results obtained include a total of 31 different apps available either on the Android and iOS operating systems. Among that set, only 7 of them are mentioned in some clinical study. It is a low number given the amount of applications available, the estimated number of active users and taking into account that kind of information is completely unknown for the digital stores’ users.

INTRODUCTION

The sleep disorders are a variety of conditions that affect the sleep quality, timing or duration of the sleep. There are more than 100 specific sleep disorders and the most frequent are insomnia, apnea and restless legs syndrome (Sleep Foundation, 2020). In the last decade, the prevalence of these disorders seem to be increasing considerably among the population (Acquavella et al., 2020); but during the last year, marked by the covid-19 pandemic, sleep problems have been widely studied and all authors agree with the fact that they have increased. Hung and Zhao (2020) reported that the 18% of the Chinese population studied had a poor quality of sleep, Stanton et al. (2020) informed about 40,7% in Australia, and Martínez-Lezaun et al. (2020) found that the 70% of the Spanish university students had, during the lockdown, worse sleep quality.

At least one of those signs characterizes these disorders: trouble to fall asleep, difficulties to stay awake during the day, imbalances in the sleep schedule or unusual behaviors that disrupt sleep (Sleep Foundation, 2020). The sleep disorders have also several consequences as memory and concentration difficulties, appetite disturbances and difficulties to perform activities of daily living (Ram, Seirawan, Kumar and Clark, 2010).

In order to improve the sleep quality, the Society of Behavioral Sleep Medicine (Crew et al., 2020) recommend maintaining certain level of physical activity and keep the exposure to natural light, since the lack of it induce mood disorders, alter the energy levels, provokes appetite disturbances and changes the sleep routine (Wright et al., 2013).

On the other hand, technological solutions have recently appeared claiming to help improve health. In this way, the term e-health was defined at the beginning of the century as an emerging field, referring to health services, characterized by a technical development but also a commitment to improve health using communication technology (Eysenbach, 2001). Related with this term, the mobile Health (m-health) refers to “medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, personal digital assistants (PDAs), and other wireless devices” (WHO, 2011).

The World Health Organization (WHO) mHealth Technical Evidence Review Group developed, in 2016, the “mHealth evidence reporting and assessment (mERA) checklist”, with the aim of improving the completeness of reporting of mobile health (mHealth) interventions (Agarwal et al., 2016).

The creation of this checklist is a starting point to try to standardize the interventions through these devices, and it certainly helps health professionals who use them, to define the content, the context and the technical features. However, this may not be enough, and before deciding to implement the use of this kind of technology, it would be useful to know if these mobile applications have or not have effect on health. Before implementing a new treatment for any health condition, it has been subjected to numerous studies that support its effectiveness and results. However, this is not widespread in terms of the use of mobile applications for health-related purposes, which raises ethical questions that are important to reflect on.

The objective of this work was to analyze, under an ethical judgement, the most recommended mobile applications focused on sleep disorders, and reflex on the ethical conflicts generated and the ethical judgment of the users.

METHODOLOGY

As the first step of this study, a search of the existing literature related to mobile applications with an effect on users' sleep was conducted. Since there is no validated scientific search method related to applications (Linares-del Rey et al., 2019) and there are a large number of applications; first, a search for mobile applications related to sleep was carried out in the main mobile application e-shops’ for the main operating systems: Android (Google Play) and iOS (App Store) search engines. In parallel, a search was carried out in the main web search portals (Google and DuckDuckGo), including the keywords "sleep mobile app" and selecting only web contents belonging to some medical, governmental or at least non-commercial association or agency, both English and Spanish. Apps that are not intended to improve sleep quality (i.e., only measure sleep time/quality) were excluded.

Subsequently, a bibliographic search of published scientific articles focused on the design, development and validation of the mobile applications that are recommended by the web content of pages detailed above was carried out. This search was performed in the following databases: PudMed and Web of Science.

As a result, only mobile applications that, at the same time, comply with all three conditions:

a) claim to have some effect on users' sleep,

b) are recommended by any content from non-commercial websites and

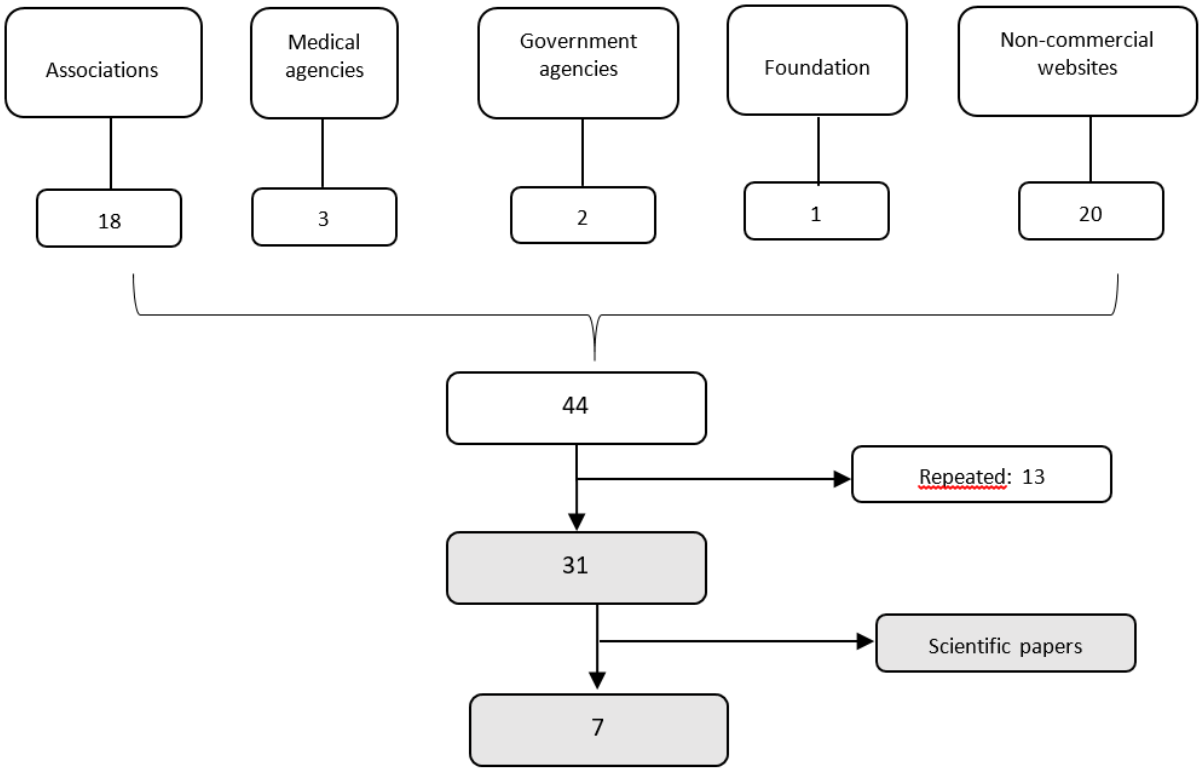
c) have been part of a scientific design, development or validation study

were included in the study.

RESULTS

The present study included 31 mobile applications that were selected from 10 websites: four of which were websites of sleep-related associations, such as the American Sleep Association or the Indian Society for Sleep Research; three were non-commercial websites, two of which were medical agencies (Asociación Argentina de Medicina del Sueño, National Health Service); one government agency (US Department for Veteran Affairs) and one foundation (Sleep Foundation). After eliminating 13 repeated applications, a search was carried out in health science databases (PudMed and Web of Science), obtaining only 7 applications that had carried out a scientific study, whether validation, design or development (see Figure 1).

Figure 1. Sleep mobile apps search and selection process.



Source: self-elaboration

Regarding the 31 apps found (Table 1), most of them were available for Android and iOS devices (a total of 17 apps/ 54%), 6 only for Android devices (19,35%) 5 only for iOS devices (16,1%) and 3 for Android, iOS and Windows devices (9,3%) . The selected apps had been updated quite recently: in 2021 (15 apps/ 48.4%), 2020 (4 apps/13%) and 2019 (6 apps/19,3%). The resources the apps use to promote sleep were very varied, but the most used was the accompanying music (10 applications/32,2%), followed by white sound, understood as sound of rain, ocean, birds, rivers, among others (7 applications/22,5%) and meditation techniques (6 applications/19,3%).

The cost of the different applications is also very varied, only 7 are free, the rest have a cost ranging from 0.50 cents to 349.99 euros. In addition, some applications set the cost depending on the elements used by the user. Most of the applications are privately financed as only two applications are financed by governmental bodies. Regarding the number of downloads of the applications according to the download search engines, 2 applications reached 10 million downloads (6,5%); 4 reached 5 million (13%); 2 reached one million (6,5%); 3 reached five hundred thousand (9,3%); 5 reached one hundred thousand (16,1%) and 9 reached ten thousand (29%).

As can be seen in Figure 1, of the 31 applications found, only 7 have been described in scientific articles published in health science databases (PudMed and Web of Science); either as a result of a validation, design or development study which represents only the 22,5% of the selected apps.

Table 1. Descriptive data of the mobile sleep apps included in the study.

App Name	Last update	Resources	Price	Downloads	Ratings
Relax & Rest Guided Meditations App [a, e]	2019	Meditation	2,09€	10.000+	484 ratings; 4,5/5
Make You Sleepy [a, e]	2019	Accompanying music	Free	10.000+	36 ratings; 2,5/5
Sleep well [b, e]	2020	Hypnosis	3,09€	10.000+	2.521 ratings; 4,5/5
White noise [b, g]	2021	White noise	1,09€	10.000+	16.820 ratings; 4,7/5
CBT-i Coach [b, e]	2019	Cognitive behavioural therapy	Free	10.000+	153 ratings; 3,8/5
Insomnia Coach [b, e]	2019	Tips, Sleep journal	Free	5.000+	16 ratings; 3,5/5
My oasis [d, e]	2021	Accompanying music	0,50€-50,00€	5.000.000+	706.709 ratings; 4,6/5
Juego fácil de dormir [c, i]	2020	Accompanying music	Free	50.000+	296 ratings; 1,9/5
Sweet Dreams [c, e]	2018	Sheep counting with relaxing music	1,19€	10000+	274 ratings; 3/5
Calm [b, f]	2021	Meditation; Sleep stories; Relaxing music	0,77€-349,99€	10.000.000+	359.668 ratings; 4/5
HeadSpace [b, f]	2021	Meditation	6,99€-129,99€	10.000.000+	207.344 ratings; 4,6/5
Ten Percent Happier [b, e]	2021	Video and Meditation	4,99€-104,99€	500.000+	13.215 ratings; 4,8/5
MyLife Meditation: Meditate, Relax & Sleep Better [b, e]	2021	Meditation	9,99€-269,99€	1.000.000+	24.673 ratings; 4,6/5
Buddhify [b, e]	2021	Meditation	4,09€	100.000+	3.284 ratings; 3,9/5
Relax & Sleep [c, e]	2017	White noise	1,99€	5.000.000+	38.626 ratings; 3,8/5
Sleep Bug [a, e]	2015	White noise	1,99€	100.000+	1.576 ratings; 4,3/5
TaoMix 2: Sleep Sounds & Focus [b, f]	2021	White noise	0,99€-6,49€	100.000+	3.168 ratings; 4,6/5
Insight Timer [b, h]	2021	Music; Stories; Meditation;	2€-60€	5.000.000+	110.340 ratings; 5/5
Relaxing Music: Sleep Sounds [b, h]	2021	Accompanying music	2€-40€	1.000.000+	14.976 ratings; 4/5
Relaxing Music to Sleep [c, i]	2020	Accompanying music	Free	100.000+	4.706 ratings; 4/5
Relax Melodies [b, h]	2021	Accompanying music	19€ a 60€	10.000.000+	30.5192 ratings; 4,5/5
White Noise Generator [b, g]	2021	White noise	Free	1.000.000+	58.555 ratings; 4,3/5
myNoise [b, e]	2019	White noise	1€-12€	100.000+	2.554 ratings; 4/5
Awoken - Lucid Dreaming Tool [c, e]	2018	Guided Conversations	0,99€-6,49€	500.000 +	No data available
Sleep Cycle-Smart Alarm Clock [b, e]	2021	Sleep journal	0,89€-64,99 €	5.000.000+	130.073 ratings; 4,7/5
Relax Melodies [b, h]	2021	Accompanying music	2,99€-329,99€	10.000.000+	305.270 ratings; 4,6/5
Sleep Cycle-Power Nap [d, e]	2014	White noise; sound landscapes	2$	No data available	657 ratings; 4/5

Pzizz-Sleep, Nap, Focus [d, e]	2018	High Frequency Sounds	10-70	No data available	83 ratings; 4,7/5
White Noise Lite [d, g]	2021	Hypnosis	Free	No data available	128.700 ratings; 4,8/5
Relax&Sleep Well-Hypnosis and Meditation [c, e]	2020	Meditation	1,19€-12,99€	500.000+	9.180 ratings; 4,7/5
Sleepio [d, e]	2021	Cognitive behavioural therapy	Free	No data available	87 ratings; 2,6/5

Operating Systems: [a]: android-IOS-Windows; [b]: android-IOS; [c]: android; [d]: IOS
Languages: e: English; f: English-Spanish; g: English-Spanish-French; h: English-Spanish-Other; i: Spanish

Source: self-elaboration

Table 2 shows the main characteristics of those apps that were included in any scientific publication. Two of them have been recommended by two different information sources, that is, non-commercial organizations; and most of them have been developed without any specific targeted population in mind. Financing of the implementation and exploitation is private for all, except CBT i-Coach, which is financed by a governmental agency. In total, the 7 applications have been included in 24 scientific publications; but 3 of them account for the 66.66% of the published reported results: Calm (6 publications), HeadSpace (5 publications) and CBT i-Coach (5 publications).

Table 2.

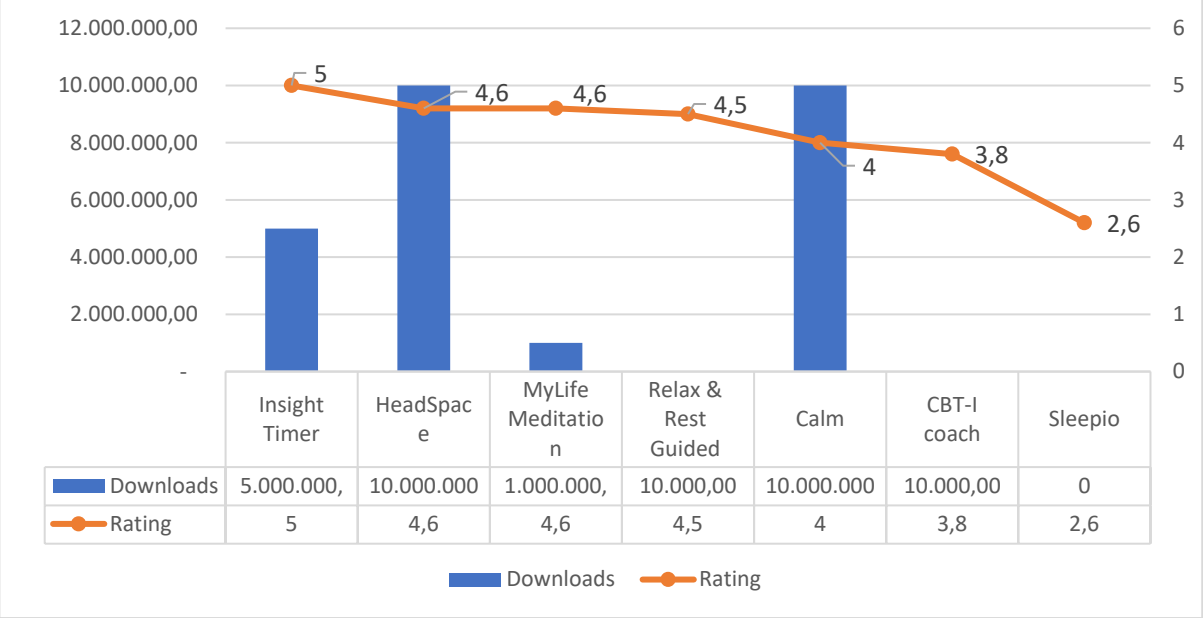
App Name	Web	Participants	Developers	Studies
Relax & Rest Guided Meditations App	Consumer Reports	ned	Meditation Oasis.	2
CBT-i Coach	Sleep Foundation Indian Society for Sleep Research	Insomniac patients	US department of veteran's affairs	5
Calm	Bestmattresses.com Sleepstation	Undefined	Calm.com Inc.	6
HeadSpace	American Sleep Association Prevention.com	Undefined	Headspace Inc.	5
MyLife Meditation: Meditate, Relax & Sleep Better	Consumer Reports	Undefined	Stop, Breathe & Think	3
Insight Timer	Prevention.com	Undefined	Insight Network Inc	1
Sleepio	National Health Service	Adults	Big health ltd	2

Source: self-elaboration

Figure 2 shows the data obtained publicly from the different on-line stores where they are officially available for their download and use, for the 7 mentioned apps. Regarding the number of downloads and average rating, it can be observed that 4 applications have an average rating between 5 and 4,5 points, while the number of downloads is higher than 10 million for the 2 most popular apps and that 4 out of 7 have reached a number of downloads of 1 million or more.

All data has been collected from the Google Play Store. Since the Sleepio app is only available on the Apple Store, authors don't have access to that information, since it is not publicly included in this platform.

Figure 2. Seven apps that are included in scientific research, comparing number of downloads and average user rating for each one.



Source: self-elaboration

DISCUSSION AND CONCLUSIONS

Europe has gone into the digital market twenty years ago; during this time, technological improvements have gone noiseless growing up, and, carving up a new knowledgeable generation. Furthermore, TIC´s mean a new way of speaking, a new language, new symbols and meaning, new tones and icons, and, consequently, some lacks right understanding about the rules and values of managing. This implies mistakes and controversies, moral dilemmas and statements about the right or wrong use of this trending usability of TIC´s and the behavior caused through its usability.

One of the most disruptive App´s, what is, Serious Games, are used by the population (youngest specifically) with a natural and innate talent, and, this is because they manage more quantity of data, information, knowledge and contents immediately, and, because the digital market offers a huge range of possibilities. As consequence, this new products paradigm force to analyze and value controversial parameters, such as, ethics, security and privacy.

This research –as a reflex on the ethical conflicts generated in the usability of serious games for improving sleep quality- reports a vague relationship between the vast list of technological applications developed and the rigorous scientific validation procedure. Despite what we might expect, just a mere 22% out of the total amount of applications can be potentially distinguish in scientific nomenclature (clinic validation). The underlying point is that TIC´s community should reflect the results of the Apps sorting. Research efforts should also focus on mainly ethics, security and privacy. Ethics, because is mandatory to research about the conflicts within the validation of the licenses; security and privacy because in the Big Data virtual world is compulsory to respect rules and regulations and guarantee products and services' know-how.

Taking into consideration another dimension of the use of technologies, what are the users themselves, it is mandatory to opened up some questions regarding their ethical judgment such as perceive risk of use (Khaled and Faqih, 2016; Littler and Melanthiou, 2006; Wiegard and Breitner, 2017), behavioral intention (Balakrishnan and Griffiths, 2018; Lee, Chung and Lee, 2013), sustainable

consumption (Mulcahy, Russell-Bennett and Iacobucci, 2020) and performance expectancy (Orji, Mandryk and Vassileva, 2017).

A key contribution is the connection with the behavioral intention of use. The number of downloads generates a feeling of adherence; potential users perceive more interest and security and the product aim to become more and more attractive. Findings point out that a mere 22.5% of the Apps have carried out some type of research study. Specifically, up to 24 research studies reinforce these Apps in their objective of improving the quality of sleep with their usability parameters. In addition, of these, more than half have 10,000,000 downloads.

The average valuation is very high, which means that there is a market very receptive to this type of products to improve rest; even those applications with hardly any downloads have some kind of medium-high positive rating. It can be inferred that the reason for this high assessment lies in the importance and needing that is given to the quality of sleep, considering that there are more and more stimuli that hinder good rest.

The study carried out also shows that it would be very interesting for each App to get the most specific general rating based on the parameters on which its research studies are based (rest time, etc.) including publications with an impact index that endorse the improvements obtained. In parallel, comments on the usability and results of the tools that are publicly exposed are also part of the instrumental value of the product. The approach in this sense cannot be merely economic, neither self-interests, it must also be ethical and based on core values (real information, legal and social policies, rational and impartial).

Based on other collected monitored data, App´s review ratings related to emotional perceptions -influencing aspects such as behavioral intent and performance experience-, therefore, moral discourse in this regard should retrieved. In this sense, outcomes suggest the importance of embedding context with real track data set by dismissing those that does not serve for the App´s purpose rather than confuse.

When commenting on the criticism form the sustainable consumption, the research does not bring new light on the issue. The ordinary discourse focused on the real taxonomy of the concept implies a responsible acquisition, responding to basic needs and optimizing quality of life. Findings point out that while in most of the Apps analyzed there is no specific target population neither premises, the use and benefits of the applications on improving sleep quality cannot be compared. So, in short, the products are susceptible of being severely hindered, because, on the one side, the users are unable to distinguish which App provides assistive in their sleeping troubles and, on the other hand, Apps turn into profoundly impersonal.

Regarding the concept of security in ICT, it is positively valued that the Apps shows a privacy policy notice. It has been verified that the rigor of security and privacy is detailed in the collection of data provided by the users, other data on the use of services (data of use, data of operations, data of registry, data of device, information generated, cookies, data from other sources), use of data, data communication, advertising and analytical services provided by others, transfer of personal data, account data, promotional communications, alerts and other notifications, legal basis, data retention and requests of the interested parties. Finally, it should be noted that it cannot be concluded that the evaluations are better or worse depending on whether the Apps is within a scientific study compared to the Apps that are not. Overall, the current research study recognizes that, under moral considerations, Serious Games -as trending digital products in the store-, are potentially debatable in terms of ethical conflicts generated by the App as a technological resource, and ethical judgment of the users. In this sense, discussion derived from the literature review, point out the opportunity this new innovative products to primarily work together, offer data security and data protection.

Technology products developed for healthcare in daily routines (quality of sleeping specifically) need to be launch in the market by testing acceptance, perceived advantages, privacy risk, limitations and benefits. A big challenge issue for the industry that build-in trusting products to reinforce its usability and feasibility.

KEYWORDS: App, Serious Games, Sleep Disorders, Ethics and Morality, Sustainable Consumption.

REFERENCES

Acquavella, J., Mehra, R., Bron, M., et al. (2020). Prevalence of narcolepsy, other sleep disorders, and diagnostic tests from 2013-2016: insured patients actively seeking care. *Journal of Clinical Sleep Medicine*. Published online August 15, 2020. https://doi.org/10.5664/jcsm.8482

Agarwal, S., LeFevre, A.E., Lee, J., L'Engle, K., Mehl, G., Sinha, C., & Labrique, A. Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. *British Medical Journal* 2016;352:i1174. https://doi.org/10.1136/bmj.i1174

Balakrishnan, J., 6 Griffiths, M.D. (2018). Loyalty towards online games, gaming addiction, and purchase intention towards online mobile in-game features. *Computers in Human Behavior*, 87, 238-246.

Crew, E.C., et al. (2020). The Society of Behavioral Sleep Medicine (SBSM) COVID-19 Task Force: Objectives and Summary Recommendations for Managing Sleep during a Pandemic. *Behavioral* Sle*ep Medicine* [Internet]. Vol 18(4), 1-3. https://doi.org/10.1080/15402002.2020.1776288

Eysenbach, G. (2001). What is e-health? *Journal of Medicine Internet Research*, 3(2): e20. https://doi.org/10.2196/jmir.3.2.e20

Khaled, M.S., & Faqih, J. (2016). An empirical analysis of factors predicting the behavioral intention to adopt Internet shopping technology among non-shoppers in a developing country context: Does gender matter? *Journal of Retailing and Consumer Services* 30, 140-164.

Lee, H.-G., Chung, S., & Lee, W.-H. (2013). Presence in virtual golf simulators: The effects of presence on perceived enjoyment, perceived value, and behavioral intention. *New Media and Society*, 15(6) 930-946.

Linares-del Rey, M., Vela-Desojo, L., & Cano-de la Cuerda, R. (2019). Aplicaciones móviles en la enfermedad de Parkinson: una revisión sistemática. Neurología, 34(1), 38-54. https://doi.org/10.1016/j.nrl.2017.03.006

Littler, D., & Melanthiou, D. (2006). Consumer perceptions of risk and uncertainty and the implications for behavior towards innovative retail services: The case of Internet Banking. *Journal of Retailing and Consumer Services*, 13(6), 431-443.

Martínez-Lezaun, I., Santamaría-Vázquez, M., & Del Líbano, M. (2020). Influence of Confinement by COVID-19 on the Quality of Sleep and the Interests of University Students. *Nature and Science of Sleep*, 12, 1075-1081. https://doi.org/10.2147/NSS.S280892

Mulcahy, R., Russell-Bennett, R., & Iacobucci, D. (2020). Designing gamified apps for sustainable consumption: A field study. *Journal of Business Research*, 106, 377-387.

Orji, R., Mandryk, L. R. & Vassileva, J. (2017). Improving the efficacy of games for change using personalization models. *ACM Transactions on Computer-Human Interaction*. 24(5), Article Number 32.

Ram, S, Seirawan, H., Kumar, S.K., & Clark, G.T. (2010). Prevalence and impact of sleep disorders and sleep habits in the United States. *Sleep and breathing*, 14(1), 63-70.

Sleep Foundation (15th December 2020) *Sleep Disorders.* [Access 15/12/2020] https://www.sleepfoundation.org/sleep-disorders

Stanton, R., et al. (2020). Depression, anxiety and stress during COVID-19: Associations with changes in physical activity, sleep, tobacco and alcohol use in Australian adults. *International Journal Environmental Research of Public Health*. June 1, 2020;17(11), 1-13.

Wiegard, R.-B., & Breitner, M.H. (2017). Smart services in healthcare: A risk-benefit-analysis of pay-as-you-live services from customer perspective in Germany. *Electronic Markets* (article in press), 1-17.

World Health Organization. mHealth (2011). New Horizons for Health through Mobile Technologies: Based on the Findings of the Second Global Survey on eHealth (Global Observatory for eHealth Series, Volume 3). http://www.who.int/goe/publications/goe_mhealth_web.pdf [accessed 2021-04-15].

Wright, K.P., et al. (2013). Entrainment of the human circadian clock to the natural light-dark cycle. *Current Biology,* 23(16), 1554–1558.

CYBORG ACCEPTANCE IN HEALTHCARE SERVICES: THE USE OF CYBORG AS A SURGEON

Ala Almahameed, Mario Arias-Oliva, Jorge Pelegrín-Borondo

Social and Business Research Lab, Universitat Rovira i Virgili (Spain), Complutense University of Madrid (Spain), University of La Rioja (Spain)

ala.almahameed@estudiants.urv.cat; mario.arias@ucm.es; jorge.pelegrin@unirioja.es

ABSTRACT

The acceptance of using cyborg technology, which is a result of combining the human biological body with insideables or/and wearables technologies, is still under investigation, and the acceptance of the services that could be offered by cyborg itself hasn't been investigated yet. This research focuses on the factors that could impact cyborg acceptance in healthcare services. In particular, a model was developed to investigate the intention to choose cyborg as a surgeon to correct the visual impairment in one eye. The PLS-SEM technique was used to test the proposed hypotheses. The proposed model's explanatory power concerning the intention to choose cyborg surgeon is high ($R^2 = 0.77$). The results confirmed the impact of effort expectancy, performance expectancy, social influence, and arousal emotional dimension on the intention to choose cyborg services. In contrast, pleasure emotional dimension, empathy, and perceived risk were not found to have any significant impact on the intention to choose the proposed cyborg surgeon. Further research is required to test the proposed model in different countries and different service settings.

INTRODUCTION

The emergence of technological implants for therapy and improvement of human capabilities opens a new era in human-machine interaction. The term cyborg (Cybernetic Organism) is introducing the human with new capabilities, by using wearables or by implanting electronic devices into humans' body (Enno Park, 2014). Most of the developed implantable devices across the past decade have been utilized for the healthcare applications, such as paralyzed limbs control, pacemakers, and cochlear (Raatikainen et al., 2015), and some of these devices are being used to enhance human capabilities, such as memory, vision, hearing, physical strength, and moral enhancement (Reinares-Lara et al., 2016). For instance, radio frequency identification (RFID) chips can be implanted under human skin to be used for access control, personal identification, credit card, and mobile payments by using near-field communication (NFC) technology (Adam & Wilkes, 2016). Furthermore, the cochlear implants (CI) represent the first interaction between the human brain and the machines to replace the lost sounds by allowing the brain to recover the sense of hearing. Also, it could be used to enhance the hearing ability of healthy people (Christie & Bloustien, 2010). On the other hand, technological tattoos, fitness trackers, smartwatches, and smart glasses are representing some examples of wearables technology (Firger, 2015). This development requires to investigate customer behavior toward such technologies. In this sense, the use of technological implants in therapy application has been accepted by society and the use of them to improve human capabilities has been partially accepted. Further investigations are ongoing to be able to understand the factors that could stimulate the acceptance of such technologies (Pelegrín-Borondo et al., 2018).

Eventually, these developments in wearables and insideables to create the cyborg are pointing out to an important concern about how the interaction would be between biological bodies and technological devices, the information processing caused by this interaction, and the impact of this interaction on the environment (Greiner, 2014). Moreover, not much is known about the moral attitude of people toward the ratio between risk and benefits of using such technology and about their preferences, expectations, and needs. Meanwhile, the acceptance may be shifted from a positive to a negative state, as the use of cyborg technologies will be shifted from therapy to enhancement purposes. For instance, CI could be considered a therapy device if the user is deaf. If not, it would be considered an enhancement. The successes of these technologies will depend on the offered benefits and people's perception of these benefits (Schicktanz et al., 2015). In parallel, once these entities (i.e., Cyborgs) become realistic and as proposed, how will humans perceive cyborg individuals in their society? Are they going to accept their existence? Are they going to interact with them normally? And suppose that cyborg will become an employee in any service setting, are people willing to accept the services offered by cyborg? Could they prefer it over human services, for instance? Accordingly, the research aims to investigate the factors that influence cyborg acceptance as an entity in society, especially in healthcare services and as a surgeon. Thus, the research developed a theoretical model to investigate the intention to choose cyborg services in the healthcare sector, considering the technological and human sides of the proposed cyborg. This model will open a new line of research in this context and will be a starting point for the researchers who are interested in studying this domain.

LITERATURE REVIEW AND HYPOTHESIS

In terms of technology, the enhancement could be visible (wearables) or invisible (Implants). Moreover, it could be organic, mechanical, or a combination of both of them. In Fact, using technological implants to create a cyborg will keep the enhanced human to look like a normal one. This means and from the appearance perspective, the enhanced humans will avoid the negative social response that could be associated with their abnormal looks (West, 2016). Consequently, the acceptance of cyborg could be unlike the machines' acceptance (e.g., robot), because the cyborg is still a human with enhanced capabilities that are beyond the normality. However, to make a complete picture, it is necessary to study cyborg acceptance from both perspectives: as a machine and as a human, since in both cases (i.e., implants and wearables), technology is the major part of the enhanced human's body.

Even though cyborg technology is still in its development stage, there are some examples and attempts to implement cyborg technology. For instance, "Neil Harbisson" has color blindness. Neil now can hear the colors through a camera placed on the front of his face. The camera captures the colors as visual signals and sends them to a chip located on the back of Neil's head. Then the chip converts the visual signals into sound waves. And through these sounds, his brain can distinguish between different colors. This "Eyeborg" gives Neil the ability to recognize colors that can be perceived by normal humans and the colors that laid beyond human vision ability. Neil is considered the first official cyborg because his Eyeborg is shown in his passport photo (Parkhurst, 2012). In fact, Neil's journey was not easy with that Eyeborg. He mentioned through an interview with BBC News that two police officers attacked him when he was visiting Paris. They thought he was filming them, and even after he told them it is for hearing sounds, they thought he was laughing at them and they crashed his Eyeborg (BBC, 2012). Actually, the literature showed how technological awareness can reduce the possibility of rejecting new technologies (Mutahar et al., 2018). In this context, special programs and campaigns could be required to increase public awareness regarding the new technologies in terms of their potential benefits for humanity (Kardooni et al., 2016).

The acceptance of new technologies includes some theories about technology acceptance, such as the Technology Acceptance Model (TAM1) for Davis (1985) and its extensions TAM2 (Venkatesh & Davis, 2000) and TAM3 (Venkatesh & Bala, 2008), the Unified Theory of Acceptance and Use of Technology (UTAUT1) for Venkatesh et al. (2003) and its extension UTAUT2 Venkatesh et al. (2012) and the Cognitive-Affective-Normative Model (CAN) for Pelegrín-Borondo et al. (2016), which has been developed based on the TAM and UTAUT models, to study the acceptance of being a cyborg. The performance expectancy, which is one of the UTAT constructs, is related to the individuals' beliefs about the system's ability to improve their job performance. And effort expectancy is related to the simplicity of using the system (Venkatesh et al., 2003). Human needs to perceive the usefulness of cyborg in terms of its superiority in performance if compared to humans. It can stimulate the acceptance of dealing with cyborgs if the consumers find it better than the other options or stimulate the rejections if there are no differences in terms of performance and outcomes. But it is important also to consider the possibility of the low effect of these two constructs in the initial investigation of cyborg acceptance since the technology is still in its novelty stage (Pelegrín-Borondo, Reinares-Lara, et al., 2017).

Since individuals are members of their social entities, other member's opinions and advice toward any behavior or decision could make a difference and could direct that behavior or decision. Therefore, it makes sense to investigate the effect of social influence while studying the acceptance of new technology (Ajzen, 1991). In fact, this side was studied in technology acceptance literature and it is one of the main constructs of technology acceptance models. The social influence was introduced by the Theory of Reasoned Action (TRA) for Fishbein and Ajzen (1975) and the Theory of Planned Behavior (TPB) for Ajzen (1991). For example, the literature showed the significant impact of this construct on the acceptance of new technologies (Davis, 1989; Venkatesh, 2000). As well, its impact on the acceptance of Nanoimplants (Pelegrín-Borondo et al., 2015, 2016; Pelegrín-Borondo, Reinares-Lara, et al., 2017; Reinares-Lara et al., 2016, 2018), breast augmentation for young women (Moser & Aiken, 2011) and on the acceptance of virtual customer integration (Füller et al., 2010).

Emotions have been considered as a way to distinguish humans from objects and machines. Furthermore, the ability to express basic emotions could be proof of humanity (Heisele et al., 2002). Moreover, as the proposed relation between humans and cyborgs will involve a direct interaction, it is essential to investigate the impact of anxiety emotion on the interaction. Indeed, the expected anxiety is a reflection of the abnormality and superpower associated with cyborg technology. Factually, the anxiety problem is not related to the technology itself, rather than it is an emergence of this negative feeling while interacting with it (Oh et al., 2017). However, changing the attention toward the benefits of the technology could help in reducing the associated anxiety during the interaction process (Reinares-Lara et al., 2016). Meanwhile, some studies claimed that anxiety is not a significant determinant of the intention toward new technologies (Pelegrín-Borondo, Reinares-Lara, et al., 2017; Venkatesh et al., 2003). In the same context, Pelegrín-Borondo et al. (2016) used emotional dimensions: positive and negative emotions in the CAN model to investigate the acceptance of being a cyborg. However, there is some degree of consensus that the arousal and pleasure emotional dimensions are the most adequate dimensions to analyze the emotional response of an individual to a stimulus (Pelegrín-Borondo et al., 2015). The level of emotional pleasure and emotional arousal are the most supported emotional dimensions by literature (Cohen, Pham, & Andrade, 2008; Pelegrín-Borondo et al., 2015; Russell, 1980, 2003). In this sense, Mehrabian and Russell (1974) and Russell and Mehrabian (1977) suggested that you can measure what a person is feeling by employing a limited number of emotional dimensions. They proposed a scale with three dimensions: pleasure, arousal, and dominance (PAD). Eroglu, Machleit, and Davis (2001) recommended using arousal and pleasure only, and without the dominance dimension. They claimed that these two dimensions can represent the

range of emotions that emerged in response to environmental stimuli and based on Russell's (1979) recommendation. Pleasure is related to a person's state of feeling of goodness, happiness, joyfulness, or contentedness in a certain situation. And, arousal is about a person's state of feeling with excitement, alert, stimulation, wakefulness, or activeness in a certain situation (Das, 2013; Mehrabian & Russell, 1974). Positive arousal and pleasure emotions can allow humans to feel optimism while choosing their plans and goals. In fact, arousal could be seen as preparation for actions (Russell, 2003). Also, pleasure can affect customer behavior toward a successful choice of a specific service or/and product. Moreover, while using a specific service, customers may develop positive or negative emotions. The positive ones are important for the future behavior of customers (Pappas et al., 2013). Furthermore, they are considered important in directing the attitude of customers toward new technologies, and they can enhance the predictive power of the technology acceptance models (Kulviwat et al., 2007).

Perceived risk (PR) was introduced by Bauer (1960) to the marketing research. It is related to the consumer perception of the uncertainty and adverse outcomes associated with buying a product or a service (Dowling & Staelin, 1994). Pavlou (2003) integrated the PR into the TAM model while studying the acceptance of e-commerce. The research results confirmed the direct impact of the PR on the behavioral intention and use behavior. In the cyborg technology context, using PR in studying the acceptance of such technology could be justified, as the technology is still under development and not much is known about it. For instance, Gao et al. (2015) pointed to the significant negative impact of PR on the intention of using wearable technologies. However, Pelegrin-Borondo et al. (2017) found PR impact on the acceptance of insideable technologies higher than its impact on the wearable ones. Moreover, when benefits exceed the risk that is associated with nanotechnologies, the perception of risk may decrease (Gupta, Fischer, & Frewer, 2015; Satterfield, Kandlikar, Beaudrie, Conti, & Harthorn, 2009).

Empathy can be seen as the degree of caring and attention that employees show to their consumers (Parasuraman et al., 1988), and it has a direct impact on consumers' positive expectations toward service quality (Bebko, 2000). In addition to its role in establishing a successful consumer-employee interaction (Homburg, Wieseke, & Bornemann, 2009). Besides, it is related to understanding consumers' perspectives and interacting with them emotionally (Davis, 1983). In fact, empathy isn't a personal treat as much as it is a skill that can be created and developed to enhance consumer-employee interaction, which may lead to consumer satisfaction (Malle & Pearce, 2001). In the same context, Fugate, Kinicki, and Ashforth (2004) defined adaptability as employees' ability and willingness to modify their feelings, thoughts, and behavior to fit consumer requirements and needs, and it is related to employee empathy too (Kieren & Tallman, 1972). Furthermore, empathy should be taken into consideration during the hiring process, as it has a major influence on the consumers' perception of service value (Namasivayam & Denizci, 2006). For instance, when a salesperson shows a high level of empathy, the consumer satisfaction level could be increased, which in turn could increase their attitude toward the offered product (Stock & Hoyer, 2005). In the technology acceptance context, such as human-robot interaction, humans can convey empathy by imitating the facial expression of the other party (Riek & Robinson, 2008). It could be proposed that this way of conveying empathy should be used in the human-cyborg interactions since the perceived empathy is a significant determinant of the intention towards humanoid technologies (Homburg & Merkle, 2019).

Based on the literature review, we posed the following hypotheses:

H1a: Patients' intention to choose the Cyborg surgeon is positively affected by effort expectancy.

H1b: Patients' intention to choose the Cyborg surgeon is positively affected by performance expectancy.

H2: Patients' intention to choose the Cyborg surgeon is positively affected by a favorable social influence.

H3a: Patients' intention to choose the Cyborg surgeon is positively affected by pleasure.

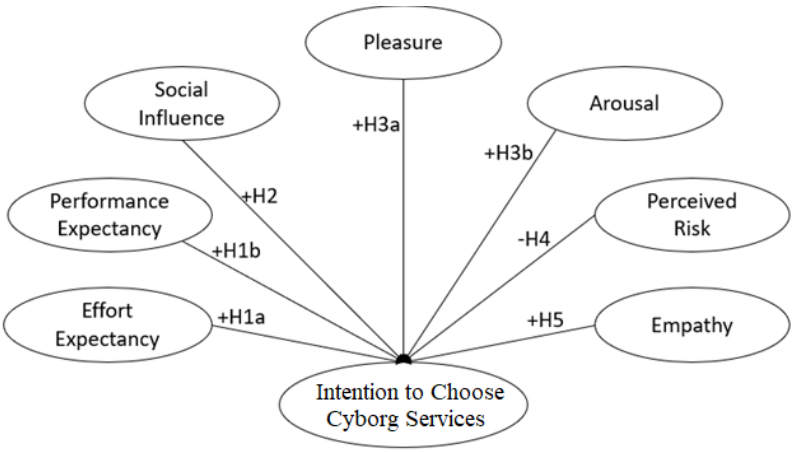
H3b: Patients' intention to choose the Cyborg surgeon is positively affected by arousal.

H4: Patients' intention to choose the Cyborg surgeon is affected negatively by the perceived risk.

H5: Patients' intention to choose the Cyborg surgeon is positively affected by perceived empathy.

A comprehensive theoretical model of variables influencing the intention to choose Cyborg services in the healthcare sector based on the proposed hypotheses is shown in Figure 1.

Figure 3. The Proposed theoretical model.



METHODOLOGY

This research used a quantitative methodology, and the online survey was developed to test research hypotheses using Google Forms. The data were collected from 379 individuals from different Jordanian universities. A total of 53% of the respondents were men, and 47% were women.

The partial least-square structural equation modeling (PLS-SEM) technique was used to examine this research model by using SmartPLS version 3 software. (Hair et al., 2011) recommend using PLS-SEM "if the goal is predicting key target constructs or identifying key 'driver' constructs," (p.144), which is the case in this research. Similarly, other authors suggest that PLS-SEM is appropriate when the research has a predictive purpose and an explanatory purpose (Henseler et al., 2016). Furthermore, PLS-SEM assesses the model relationships in a series of ordinary least squares (OLS) regressions, in order to maximize the explained variance of the endogenous latent variables. The sequence of OLS regressions makes PLS-SEM achieve a higher level of statistical power and lower demand concerning the sample size (Reinartz et al., 2009).

An 11-point scale (0 to 10) was used for the measurement scale, which was developed based on the literature review. The measurement scale developed by Venkatesh and Davis (2000) was used to measure the intention to choose cyborg services, which has been used and validated by various previous technology acceptance studies in different service settings (Chen et al., 2017; Chow et al., 2013; Im, Kim, & Han, 2007; Heijden, 2004) In contrast, the measurement scales for effort expectancy, performance expectancy, and social influence constructs were developed based on Venkatesh et al.

(2012) measurement scale, which has been used in previous studies in technology acceptance for healthcare contexts (Alaiad & Zhou, 2013, 2014; Alaiad et al., 2013; Graaf, Allouch, & Dijk, 2019; Hossain et al., 2019).

The scale that has been used to measure perceived risk was developed based on the scale adopted by Faqih (2016) which was developed by Shim, Eastlick, Lotz, and Warrington (2001) and has been validated by different studies in the technology acceptance context (Pelegrin-Borondo et al., 2017; Yang, Pang, Liu, Yen, & Michael Tarn, 2015).

Regarding the emotional dimensions of arousal and pleasure, the researchers adopted the scale developed by Mazaheri, Richard, and Laroche (2011), and used by Loureiro (2015).This scale has been used also in technology acceptance studies (Chen, Chang, & Chen, 2017; Pelegrín-Borondo, Arias-Oliva, & Olarte-Pascual, 2017; Ruiz-Mafe, Chatzipanagiotou, & Curras-Perez, 2018).

Homburg and Merkle (2019) studied attitudes toward humanoid robots and developed their measurement scale for empathy based on Davis (1983), Hogan, Hogan, and Busch (1984), and Parasuraman, Berry, and Zeithaml (1991). This scale has been used to measure empathy in this research.

The data has been collected randomly from 379 individuals in different Jordanian universities. 47% of the respondents were females and 53% were males, 75% were from the 18-30 age group, and 66% of them are in bachelor degree.

RESULTS

Measurement Model Assessment

The internal consistency reliability of the measurement model has been confirmed since the values of Cronbach's alpha and the composite reliability for all model constructs were higher than 0.70. In addition, the standardized loading of constructs indicators was greater than 0.70 and the t-values were greater than 1.96, to ensure the correct reliability indicator in the measurement model. Regarding convergent validity, all constructs have AVE values greater than 0.50, which confirmed the convergent validity of the measurement model. Table 1 shows loading values, Cronbach's alpha, and Composite reliability values. For the discriminant validity evaluation, HTMT values were less than 0.90 for all constructs, and the square root of the AVE value for each construct was higher than the correlation value with the other constructs (Table 2).

Table 1. Internal Consistency Reliability & Convergent Validity.

Variable	Indicators	Loading	Cronbach's alpha	Composite reliability	AVE
AR	AR1	0.957	0.907	0.956	.915
	AR2	0.956			
EE	EE1	0.931	0.962	0.972	0.898
	EE2	0.955			
	EE3	0.951			
	EE4	0.954			
EM	EM1	0.871	0.956	0.966	0.852
	EM2	0.897			
	EM3	0.947			
	EM4	0.958			
	EM5	0.939			
IC	IC1	0.977	0.952	0.977	0.954
	IC2	0.977			

PL	PL1	0.952	0.883	0.945	0.895
	PL2	0.940			
PR	PR1	0.970	0.931	0.943	0.848
	PR2	0.923			
	PR3	0.867			
PE	PE1	0.930	0.943	0.959	0.854
	PE2	0.940			
	PE3	0.938			
	PE4	0.886			
SI	SI1	0.948	0.942	0.963	0.896
	SI2	0.958			
	SI3	0.935			

PE: Performance Expectancy, **EE**: Effort Expectancy, **SI**: Social Influence, **PR**: Perceived Risk, **EM**: Empathy, **AR**: Arousal, **PL**: Pleasure, and **IC**: Intention to Choose.

Table 2. Discriminant validity.

Variable	AR	EE	EM	IC	PR	PE	PL	SI
AR	**0.957**	0.743	0.729	0.772	0.095	0.757	0.834	0.744
EE	0.694	**0.948**	0.864	0.876	0.069	0.889	0.729	0.843
EM	0.679	0.829	**0.923**	0.735	0.201	0.797	0.678	0.817
IC	0.717	0.840	0.702	**0.977**	0.059	0.849	0.721	0.819
PR	0.106	0.087	0.197	0.074	**0.921**	0.151	0.100	0.165
PE	0.703	0.852	0.761	0.810	0.163	**0.924**	0.686	0.880
PL	0.748	0.675	0.626	0.663	0.101	0.632	**0.946**	0.691
SI	0.688	0.803	0.776	0.776	0.189	0.831	0.633	**0.947**

Note: Bold font values in diagonal are the square roots of the AVEs below the diagonal: correlations between the constructs, and above the diagonal: HTMT values.

Structural Model Assessment

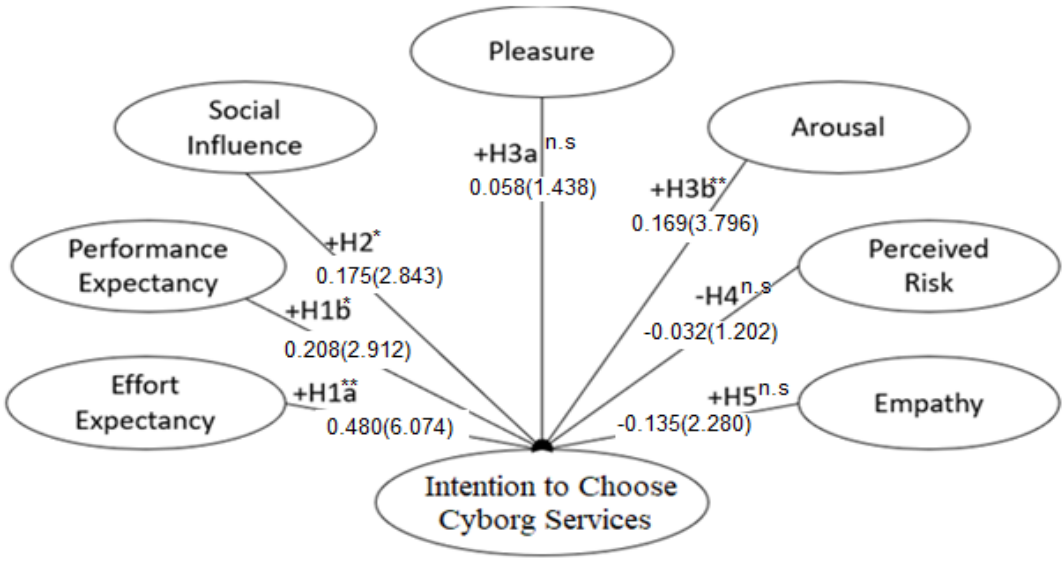
According to the research results for the structural model, H1 (The influence of Effort Expectancy and Performance Expectancy), H2 (The influence of Social Influence), and H3b (The influence of Arousal) were supported. However, H3a (The influence of Pleasure), H4 (The influence of Empathy), and H5 (The influence of Perceived Risk) were rejected, since the path coefficient wasn't significant ($0.01 < p$ and $t < 2.57$). Also, the R^2 value was 0.770, which confirmed the predictive power of the cyborg services model. Finally, Stone-Geisser's Q^2 value was 0.741, which confirmed the predictive relevance of the research model. The value of Q^2, R^2, p-value, t-value, path coefficient and support of hypotheses are shown in Table 3, and the sign, magnitude, and significance of the path coefficients are shown in Figure 2.

Table 3. Path Coefficient of Cyborg Services Hypotheses.

Variable	R^2	Q^2	Path Coefficient	t-value	Decision
Intention to Choose	0.770	0.741			
Arousal -> (+) Intention to Choose			0.169	3.796	Supported**
Effort Expectancy -> (+) Intention to Choose			0.480	6.074	Supported**
Empathy -> (+) Intention to Choose			-0.135	2.280	Not Supported
Perceived Risk -> (+) Intention to Choose			-0.032	1.202	Not Supported
Performance expectancy -> (+) Intention to Choose			0.208	2.912	Supported*
Pleasure -> (+) Intention to Choose			0.058	1.438	Not Supported
Social Influence -> (+) Intention to Choose			0.175	2.843	Supported*

Significant at P**<0.001, P*<0.01.

Figure 4. Sign, Magnitude, and Significance of the Path Coefficients. *p < 0.01; **p < 0.001; n.s = not significant.



DISCUSSION, CONCLUSION, AND IMPLICATIONS

The acceptance of using cyborg technology, which is a result of combining the human biological body with insideables or/and wearables technologies, is still under investigation, and the acceptance of the services that could be offered by cyborg itself hasn't been investigated yet. In this context, nothing is known about the moral attitude of people toward the ratio between risk and benefits of using cyborg services and about their preferences, expectations, and needs (Schicktanz et al., 2015). On the other hand, cyborgs will evolve and alter the workforce and marketplace. What is unclear is the extent of this development and its impact, and how consumers will perceive them in service settings. Thus, the research model has been developed to evaluate the patients' intention toward choosing cyborg services when compared mainly to human surgeons. The model has been built based on the technology acceptance models (e.g., UTAUT, TAM, and CAN models), in addition to integrating empathy, emotional dimensions, and perceived risk into the proposed model.

Based on PLS-SEM results, the cyborg services model can explain 77% of the variance in the intention to choose cyborg in healthcare services. This means the research model is highly predictive of the intention to choose cyborg services. So far, the inclusion of emotional dimension, perceived risk, and empathy into this research model enhanced variance explained values (R^2) when compared, for instance to the values obtained by UTAUT (44%) and CAN (73.9%) models. These results confirmed the value of extending the factors that could determine the new technology acceptance, such as emotional dimension (i.e. consumer pleasure and arousal), consumer perceived risk, and cyborg empathic behavior.

The models assessed performance expectancy, effort expectancy, social influence, empathy, perceived risk, and emotional dimension. Four of the examined variables affected the intention to choose cyborg services, except for perceived risk, empathy, and pleasure. Therefore, H1, H2, and H3b were accepted, but H3a, H4, and H5 were rejected.

According to the research results, the effort expectancy showed the most significant impact on the intention to choose cyborg services in the positive direction (H1a). Where it got the lowest p-value ($p<0.001$). Meanwhile, it has the highest t-value (6.074), which represents the highest explanatory capacity for the cyborg services model. Performance expectancy (H1b) got a t-value of 2.912, which

represents the third-highest score regarding the explanatory capacity of the research model. This is expected because many of the previous studies of cyborg technology acceptance have agreed on the importance of these variables in stimulating the intention to choose such technology. (e.g., Olarte-Pascual et al., 2015; Pelegrín-Borondo, Reinares-Lara, et al., 2017; Pelegrín-Borondo et al., 2016; Reinares-Lara et al., 2016). In addition to that, the differences between these two variables are limited to their impact on behavioral intention in terms of direction and strength of this impact. (Conti et al., 2017, 2015; Graaf et al., 2015). The importance of performance and effort expectancies could be justified because users could consider simplicity and performance efficiency as the most important factors that could stimulate their intention to choose new technologies, especially during the early stages of these technologies' emergence and use (Heerink et al., 2010, 2008, 2009).

The results showed that social influence (H2) has a positive significant impact on the intention to choose cyborg services. It got a p-value of less than 0.01, and t-value equal to 2.843. These results are in line with the previous studies about being cyborg acceptance (e.g., Olarte-Pascual et al., 2015; Pelegrin-Borondo et al., 2017; Pelegrín-Borondo, Reinares-Lara, et al., 2017; Pelegrín-Borondo et al., 2016; Reinares-Lara et al., 2018, 2016). In general, individuals could change their feelings, thoughts, attitudes, or behaviors when communicating with other individuals. Consequently, individuals could build their decisions based on other individuals' suggestions, especially when the service or product is relatively new and/or unknown (Talukder et al., 2019). Hence, the confirmed impact of social influence on the intention to choose the proposed services can justify the importance of others' advice, especially for the cyborg services, which are still in the novelty stage.

The results confirmed the impact of arousal (H3b) on the intention to choose cyborg surgeons, but it didn’t show a significant impact of pleasure emotion (H3a). Whereas, the explanatory capacity of arousal was in the second place and behind effort expectancy ($p<0.001$, $t=3.769$). In the services sector, consumers may require fulfilling their needs from two perspectives: performance and psychological perspectives. The psychological need is related to the consumer's emotions and its importance is dependent on the service nature. For instance, emotions could be considered a major criterion in hospitality services (Lu et al., 2019) and it could not impact surgeon choice (Yahanda et al., 2016). In the same context, pleasure is related to the hedonic motivation to adopt new technologies (Talukder et al., 2019). And since cyborg is a human with advanced capabilities, this could justify why it didn’t show a significant impact on intention to choose cyborg services. On the contrary, the idea of a cyborg surgeon would stimulate excitement feeling toward this future technology, which may justify the significant impact of arousal on intention to choose cyborg surgeon.

The results also didn’t confirm the negative impact of perceived risk (H4) on the intention to choose cyborg services. In fact, few studies about insideables and wearables acceptance have integrated the perceived risk into their research models. For instance, Yang et al. (2016) studied the impact of perceived risk on the intention to use wearable technology and their research results confirmed its negative impact. Contrariwise, Murata, Arias-Oliva, and Pelegrín-Borondo (2019) didn’t find a significant impact of the said construct on the acceptance to become a cyborg. In general, the inverse relation between expected benefits and risk could explain the results (Featherman, 2001; Gupta et al., 2015; Satterfield et al., 2009), because the cyborg is still a human with advanced capabilities that could be considered an opportunity to get better healthcare services, not a threat. In other words, patient perception of the cyborg benefits could reduce their perception of the associated risk while choosing a cyborg surgeon (Gupta et al., 2015). Meanwhile, the human side of the proposed cyborg surgeon could reduce the perceived risk and the uncertainty if the alternative is a technology (e.g., robots surgeon).

The result didn't show a significant impact of empathy (H5) on the intention to choose cyborg services. Actually, empathy is a skill that can be gained and developed and not a personal trait. However, in some service settings, it could be considered a significant driver of consumer purchase behavior, especially when direct interaction between employees and consumers is involved. Because the consumers in such settings expect the employee to understand their needs and to act accordingly (Malle & Pearce, 2001). Additionally, empathy has been integrated into the service quality model to investigate gaps between consumer expectations and perception of service quality (Purcarea et al., 2013). As well, in some service settings, professionalism could be considered the most important determinant of choice criteria, such as in healthcare services, which could minimize the importance of empathy on the choice decision (Wu et al., 2015). Precisely, the impact of empathy could be significant while choosing primary care physicians and psychiatrists, not the surgeons (Dehning et al., 2014; Nadi et al., 2016).

This research opens a new line of researches related to the acceptance of cyborg technology as an entity. In this regard, few studies have been conducted to investigate cyborg acceptance, which supported the companies in promoting their related products (i.e. wearables and implants) and understanding the factors stimulating their acceptance. At the same time, the acceptance of the proposed cyborg services will help the service providers to know the factors that can lead to the acceptance of hiring cyborg in a specific service setting. As a result, the developers and manufacturers of cyborg products can build their designs to match consumers' needs and based on their expectations of these enhancements. For instance, the results confirmed the impact of effort expectancy, performance expectancy, social influence, and arousal on the acceptance of the proposed cyborg services. Publicizing awareness about the simplicity in dealing and interacting with cyborgs and the superiority of their performance will be required to convince society about accepting cyborg services. In addition to that, to be served by an enhanced human could make consumers excited about the idea itself. This consequently requires reinforcing those emotions by promoting the superior capabilities of cyborg.

LIMITATIONS

One of the research limitations is related to investigate the ethical impact while studying the acceptance of such technology. Cyborg surgeons are representing an advanced technology that may have the ability to imitate and/or exceed human abilities. If these futuristic surgeons become a reality, they will compete with human surgeons and could eventually replace them. Thereby, increasing the professional and social gap between humans on one side, robots, and enhanced humans on the other side. Another ethical concern is, if these advanced surgeons are available for high-income consumers, it could create a new social class that can buy the proposed superior services. This could consequently increase the equity gap too. Furthermore, the study has been conducted in a single country. The differences in culture could affect consumers' intentions toward cyborg technology. According to that, this research should be extended to different countries for evaluating the impact of cultural differences on the intention to choose the proposed services. In addition, consumers' knowledge about cyborg technology is limited. Therefore, this research results represented a general belief of the consumers about advanced technologies. Even though the proposed services are still under the development stage, enhancing respondents' awareness about these technologies could affect their perception of the proposed services. Consequently, future research could investigate whether providing participants more information about these technologies before the data collection process - through, for instance, video demonstrations and prototypes – can impact their perception towards cyborg services and their intention to adopt them. In the same context, this research proposed a

specific use of cyborg technology. The result could vary if the proposed use is conducted in different service settings. Therefore, future research could apply this research model to different service settings.

KEYWORDS: Cyborg. Healthcare Services, Technology Acceptance, Intention to Choose.

REFERENCES

Adam, N., & Wilkes, W. (2016, September 18). *When Information Storage Gets Under Your Skin*. Wall Street Journal. https://www.wsj.com/articles/when-information-storage-gets-under-your-skin-1474251062

Agag, G. M., & El-Masry, A. A. (2017). Why do consumers trust online travel websites? Drivers and outcomes of consumer trust towards online travel websites. *Journal of Travel Research*, *56*(3), 347–369. https://doi.org/10.1177/0047287516643185

Ajzen, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Alaiad, A., & Zhou, L. (2013). Patients' Behavioral Intention Toward Using Healthcare Robots. *Proceedings of the Nineteenth Americas Conference on Information Systems*.

Alaiad, A., & Zhou, L. (2014). The Determinants of Home Healthcare Robots Adoption: An Empirical Investigation. *International Journal of Medical Informatics*, *83*(11), 825–840. https://doi.org/10.1016/j.ijmedinf.2014.07.003

Alaiad, A., Zhou, L., & Koru, G. (2013). An Empirical Study of Home Healthcare Robots Adoption Using the UTUAT Model. *Transactions of the International Conference on Health Information Technology Advancement 2013*, *2*(1), 185–198.

Bauer, R. A. (1960). Consumer behavior as risk taking. *Proceedings of the 43rd National Conference of the American Marketing Assocation, June 15, 16, 17, Chicago, Illinois, 1960*.

BBC. (2012, February 15). *The man who hears colour*. https://www.bbc.com/news/magazine-16681630

Bebko, C. P. (2000). Service Intangibility and its Impact on Consumer Expectations of Service Quality. *Journal of Services Marketing*, *14*(1), 9–26. https://doi.org/10.1108/08876040010309185

Chen, T. L., Bhattacharjee, T., Beer, J. M., Ting, L. H., Hackney, M. E., Rogers, W. A., & Kemp, C. C. (2017). Older adults' acceptance of a robot for partner dance-based exercise. *PLoS ONE*, *12*(10), 1–29. https://doi.org/10.1371/journal.pone.0182736

Chen, W.-K., Chang, D.-S., & Chen, C.-C. (2017). The Role of Utilitarian and Hedonic Values on Users' Continued Usage and Purchase Intention in a Social Commerce Environment. *Journal of Economics and Management*, *13*(2), 193–220. https://doi.org/10.1016/j.ijhm.2016.06.007

Chow, M., Chan, L., Lo, B., Chu, W. P., Chan, T., & Lai, Y. M. (2013). Exploring the Intention to Use a Clinical Imaging Portal for Enhancing Healthcare Education. *Nurse Education Today*, *33*(6), 655–662. https://doi.org/10.1016/j.nedt.2012.01.009

Christie, E., & Bloustien, G. (2010). I-cyborg: Disability, affect and public pedagogy. *Discourse: Studies in the Cultural Politics of Education*, *31*(4), 483–498. https://doi.org/10.1080/01596306.2010.504364

Cohen, J. B., Pham, M. T., & Andrade, E. B. (2008). The Nature and Role of Affect in Consumer Behavior. In *Handbook of Consumer Psychology* (1st ed., pp. 297–348). https://doi.org/10.4324/9780203809570.ch11

Conti, D., Di Nuovo, S., Buono, S., & Di Nuovo, A. (2015). A Cross-Cultural Study of Acceptance and Use of Robotics by Future Psychology Practitioners. *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication*, 555–560. https://doi.org/10.1109/ROMAN.2015.7333601

Conti, D., Di Nuovo, S., Buono, S., & Di Nuovo, A. (2017). Robots in Education and Care of Children with Developmental Disabilities: A Study on Acceptance by Experienced and Future Professionals. *International Journal of Social Robotics*, *9*(1), 51–62. https://doi.org/10.1007/s12369-016-0359-6

Das, G. (2013). The Effect of Pleasure and Arousal on Satisfaction and Word-of-Mouth: An Empirical Study of the Indian Banking Sector. *Vikalpa*, *38*(2), 95–103. https://doi.org/10.1177/0256090920130206

Davis, F. D. (1985). A technology acceptance model for empirically testing new end-user information systems: Theory and results. In *Doctoral dissertation*. Massachusetts Institute of Technology.

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease Of Use , And User Acceptance. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Davis, M. H. (1983). Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

Dehning, S., Reiß, E., Krause, D., Gasperi, S., Meyer, S., Dargel, S., Müller, N., & Siebeck, M. (2014). Empathy in High-Tech and high-touch medicine. *Patient Education and Counseling*, *95*(2), 259–264. https://doi.org/10.1016/j.pec.2014.01.013

Dowling, G. R., & Staelin, R. (1994). A Model of Perceived Risk and Intended Risk-Handling Activity. *Journal of Consumer Research*, *21*(1), 119–134. https://doi.org/10.1086/209386

Eroglu, S. A., Machleit, K. A., & Davis, L. M. (2001). Atmospheric qualities of online retailing: A conceptual model and implications. *Journal of Business Research*, *54*(2), 177–184. https://doi.org/10.1016/S0148-2963(99)00087-9

Faqih, K. M. S. (2016). An Empirical Analysis of Factors Predicting the Behavioral Intention to Adopt Internet Shopping Technology Among Non-Shoppers in a Developing Country Context: Does Gender Matter? *Journal of Retailing and Consumer Services*, *30*, 140–164. https://doi.org/10.1016/j.jretconser.2016.01.016

Featherman, M. S. (2001). Extending the Technology Acceptance Model by Inclusion of Perceived Risk. *AMCIS 2001 Proceedings*, 758–760.

Firger, J. (2015). *"Tech Tats" Usher in New Generation of Wearables*. https://www.newsweek.com/2015/12/18/tech-tats-usher-new-generation-wearables-401536.html

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Addison-Wesley.

Fugate, M., Kinicki, A. J., & Ashforth, B. E. (2004). Employability: A Psycho-Social Construct, its Dimensions, and Applications. *Journal of Vocational Behavior*, *65*(1), 14–38. https://doi.org/10.1016/j.jvb.2003.10.005

Füller, J., Faullant, R., & Matzler, K. (2010). Triggers for Virtual Customer Integration in the Development of Medical Equipment - From a Manufacturer and a User's Perspective. *Industrial Marketing Management*, *39*, 1376–1383. https://doi.org/10.1016/j.indmarman.2010.04.003

Gao, Y., Li, H., & Luo, Y. (2015). An Empirical Study of Wearable Technology Acceptance in Healthcare. *Industrial Management and Data Systems*, *115*(9), 1704–1723. https://doi.org/10.1108/IMDS-03-2015-0087

Graaf, M. M. A. de, Allouch, S. Ben, & Klamer, T. (2015). Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot. *Computers in Human Behavior*, *43*, 1–14. https://doi.org/10.1016/j.chb.2014.10.030

Graaf, M. M. A. de, Allouch, S. Ben, & van Dijk, J. A. G. M. (2019). Why Would I Use This in My Home? A Model of Domestic Social Robot Acceptance. *Human-Computer Interaction*, *34*(2), 115–173. https://doi.org/10.1080/07370024.2017.1312406

Greiner, S. (2014). Cyborg Bodies—Self-Reflections on Sensory Augmentations. *NanoEthics*, *8*(3), 299–302. https://doi.org/10.1007/s11569-014-0207-9

Gupta, N., Fischer, A. R. H., & Frewer, L. J. (2015). Ethics, Risk and Benefits Associated with Different Applications of Nanotechnology: a Comparison of Expert and Consumer Perceptions of Drivers of Societal Acceptance. *NanoEthics*, *9*(2), 93–108. https://doi.org/10.1007/s11569-015-0222-5

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a Silver Bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139–152. https://doi.org/10.2753/MTP1069-6679190202

Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Assessing Acceptance of Assistive Social Agent Technology by Older Adults: the Almere Model. *International Journal of Social Robotics*, *2*(4), 361–375. https://doi.org/10.1007/s12369-010-0068-5

Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2008). Enjoyment, Intention to Use And Actual Use of a Conversational Robot by Elderly People. *3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 113–119. https://doi.org/10.1145/1349822.1349838

Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2009). Measuring Acceptance of an Assistive Social Robot: A Suggested Toolkit. *IEEE International Workshop on Robot and Human Interactive Communication*, 528–533. https://doi.org/10.1109/ROMAN.2009.5326320

Heisele, B., Serre, T., Pontil, M., Vetter, T., & Poggio, T. (2002). Categorization by Learning and Combining Object Parts. *Advances in Neural Information Processing Systems*, *14*(2), 1239–1245.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2016). Testing Measurement Invariance of Composites Using Partial Least Squares. *International Marketing Review*, *33*(3), 405–431. https://doi.org/10.1108/IMR-09-2014-0304

Hogan, J., Hogan, R., & Busch, C. M. (1984). How to measure service orientation. *Journal of Applied Psychology*, *69*(1), 167.

Homburg, C., Wieseke, J., & Bornemann, T. (2009). Implementing the Marketing Concept at the Employee–Customer Interface: The Role of Customer Need Knowledge. *Journal of Marketing*, *73*(4), 64–81. https://doi.org/10.1509/jmkg.73.4.64

Homburg, N., & Merkle, M. (2019). A Cross-Country Comparison of Attitudes toward Humanoid Robots in Germany, the US , and India. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 4773–4782. https://doi.org/10.24251/HICSS.2019.575

Hossain, A., Quaresma, R., & Rahman, H. (2019). Investigating Factors Influencing the Physicians' Adoption of Electronic Health Record (EHR) in Healthcare System of Bangladesh: An Empirical Study. *International Journal of Information Management*, *44*, 76–87. https://doi.org/10.1016/j.ijinfomgt.2018.09.016

Im, I., Kim, Y., & Han, H. J. (2007). The Effects of Perceived Risk and Technology Type on Users' Acceptance of Technologies. *Information and Management*, *45*(1), 1–9. https://doi.org/10.1016/j.im.2007.03.005

Kardooni, R., Yusoff, S. B., & Kari, F. B. (2016). Renewable energy technology acceptance in Peninsular Malaysia. *Energy Policy*, *88*, 1–10. https://doi.org/10.1016/j.enpol.2015.10.005

Kieren, D., & Tallman, I. (1972). Spousal Adaptability: An Assesment of Marital Competence. *Journal Of Marriage And The Family*, *34*(2), 247–256.

Kulviwat, S., Bruner, G. C., Kumar, A., Nasco, S. A., & Clark, T. (2007). Toward a Unified Theory of Consumer Acceptance Technology. *Psychology & Marketing*, *24*(12), 1059–1084. https://doi.org/10.1002/mar.20196

Loureiro, S. M. C. (2015). The Role of Website Quality on PAD, Attitude and Intentions to Visit and Recommend Island Destination. *International Journal of Tourism Research*, *17*, 545–554. https://doi.org/10.1002/jtr.2022

Lu, L., Cai, R., & Gursoy, D. (2019). Developing and Validating a Service Robot Integration Willingness Scale. *International Journal of Hospitality Management*, *80*, 36–51. https://doi.org/10.1016/j.ijhm.2019.01.005

Lu, Y., Papagiannidis, S., & Alamanos, E. (2019). Exploring the emotional antecedents and outcomes of technology acceptance. *Computers in Human Behavior*, *90*(May 2018), 153–169. https://doi.org/10.1016/j.chb.2018.08.056

Malle, B. F., & Pearce, G. E. (2001). Attention to Behavioral Events During Interaction: Two Actor–Observer Gaps and Three Attempts to Close Them. *Journal of Personality and Social Psychology*, *81*(2), 278–294. https://doi.org/10.1037/0022-3514.81.2.278

Mazaheri, E., Richard, M. O., & Laroche, M. (2011). Online consumer behavior: Comparing Canadian and Chinese website visitors. *Journal of Business Research*, *64*(9), 958–965. https://doi.org/10.1016/j.jbusres.2010.11.018

Mehrabian, A., & Russell, J. A. (1974). The Basic Emotional Impact of Environments. *Perceptual and Motor Skills*, *38*(1), 283–301. https://doi.org/10.2466/pms.1974.38.1.283

Moser, S. E., & Aiken, L. S. (2011). Cognitive and Emotional Factors Associated with Elective Breast Augmentation among Young Women. *Psychology & Health*, *26*(1), 41–60. https://doi.org/10.1080/08870440903207635

Murata, K., Arias-Oliva, M., & Pelegrín-Borondo, J. (2019). Cross-Cultural Study about Cyborg Market Acceptance: Japan versus Spain. *European Research on Management and Business Economics*, *25*, 129–137. https://doi.org/10.1016/j.iedeen.2019.07.003

Mutahar, A. M., Daud, N. M., Ramayah, T., Isaac, O., & Aldholay, A. H. (2018). The effect of awareness and perceived risk on the technology acceptance model (TAM): mobile banking in Yemen. *International Journal of Services and Standards*, *12*(2), 180–204.

Nadi, A., Abedini, E., Shojaee, J., Siamian, H., Rostami, and, & Abedi, G. (2016). Patients' Expectations and Perceptions of Service Quality in the Selected Hospitals. *Medical Archives*, *70*(2), 135. https://doi.org/10.5455/medarh.2016.70.135-139

Namasivayam, K., & Denizci, B. (2006). Human Capital in Service Organizations: Identifying Value Drivers. *Journal of Intellectual Capital*, *7*(3), 381–393. https://doi.org/10.1108/14691930610681465

Oh, C., Lee, T., Kim, Y., Park, S., Kwon, S. bom, & Suh, B. (2017). Us vs. Them: Understanding Artificial Intelligence Technophobia over the Google DeepMind Challenge Match. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 2523–2534. https://doi.org/10.1145/3025453.3025539

Olarte, C., Pelegrín, J., & Reinares, E. (2017). Model of acceptance of a new type of beverage: Application to natural sparkling red wine. *Spanish Journal of Agricultural Research*, *15*(1), 1–11. https://doi.org/10.5424/sjar/2017151-10064

Olarte-Pascual, C., Pelegrín-Borondo, J., & Reinares-Lara, E. (2015). Implants to increase innate capacities: Integrated vs. apocalyptic attitudes. Is there a new market? *Universia Business Review*, *2015*(48), 86–117.

Pappas, I. O., Giannakos, M. N., & Chrissikopoulos, V. (2013). Do Privacy and Enjoyment Matter in Personalized Services? *International Journal of Digital Society*, *4*(1), 705–713. https://doi.org/10.20533/ijds.2040.2570.2013.0091

Parasuraman, A., Berry, L. L., & Zeithaml, V. A. (1991). Refinement and Reassessment of the SERVQUAL Scale. *Journal of Retailing*, *67*(4), 420–451.

Parasuraman, A. P., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. *Journal of Retailing*, *64*(1), 12–40.

Park, Enno. (2014). Ethical Issues in Cyborg Technology: Diversity and Inclusion. *NanoEthics*, *8*(3), 303–306. https://doi.org/10.1007/s11569-014-0206-x

Park, Eunil, & Pobil, A. P. del. (2013). Users' Attitudes Toward Service Robots in South Korea. *Industrial Robot: An International Journal*, *40*(1), 77–87. https://doi.org/10.1108/01439911311294273

Parkhurst, A. (2012). Becoming Cyborgian: Procrastinating the Singularity. *The New Bioethics*, *18*(1), 68–80. https://doi.org/10.1179/2050287713Z.0000000006

Pavlou, P. A. (2003). Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the Technology Acceptance Model. *International Journal of Electronic Commerce*, *7*(3), 69–103. https://doi.org/10.1080/10864415.2003.11044275

Pelegrín-Borondo, J., Arias-Oliva, M., Murata, K., & Souto-Romero, M. (2018). Does Ethical Judgment Determine the Decision to Become a Cyborg? *Journal of Business Ethics*, 1–13. https://doi.org/10.1007/s10551-018-3970-7

Pelegrín-Borondo, J., Arias-Oliva, M., & Olarte-Pascual, C. (2017). Emotions, price and quality expectations in hotel services. *Journal of Vacation Marketing*, *23*(4), 322–338. https://doi.org/10.1177/1356766716651305

Pelegrín-Borondo, J., Juaneda-Ayensa, E., González-Menorca, L., & González-Menorca, C. (2015). Dimensions and basic emotions: A complementary approach to the emotions produced to tourists by the hotel. *Journal of Vacation Marketing*, *21*(4), 351–365. https://doi.org/10.1177/1356766715580869

Pelegrin-Borondo, J., Orito, Y., Fukuta, Y., Murata, K., Arias-Oliva, M., & Adams, A. A. (2017). From a Science Fiction to the Reality: Cyborg Ethics in Japan. *ORBIT Journal*, *1*(2), 1–15. https://doi.org/10.29297/orbit.v1i2.42

Pelegrín-Borondo, J., Reinares-Lara, E., & Olarte-Pascual, C. (2017). Assessing the acceptance of technological implants (the cyborg): Evidences and challenges. *Computers in Human Behavior*, *70*, 104–112. https://doi.org/10.1016/j.chb.2016.12.063

Pelegrín-Borondo, J., Reinares-Lara, E., Olarte-Pascual, C., & Garcia-Sierra, M. (2016). Assessing the moderating effect of the end user in consumer behavior: The acceptance of technological implants to increase innate human capacities. *Frontiers in Psychology*, *7:132*, 1–13. https://doi.org/10.3389/fpsyg.2016.00132

Purcarea, V. L., Cheorghe, I. R., & Petrescu, C. M. (2013). The Assessment of Perceived Service Quality of Public Health Care Services in Romania Using the SERVQUAL Scale. *Procedia Economics and Finance*, *6*, 573–585. https://doi.org/10.1016/S2212-5671(13)00175-5

Raatikainen, M. J. P., Arnar, D. O., Zeppenfeld, K., Merino, J. L., Levya, F., Hindriks, G., & Kuck, K. H. (2015). Statistics on the Use of Cardiac Electronic Devices and Electrophysiological Procedures in the European Society of Cardiology Countries: 2014 report from the European Heart Rhythm Association. *Europace*, *17*, i1–i75. https://doi.org/10.1093/europace/euu300

Reinares-Lara, E., Olarte-Pascual, C., & Pelegrín-Borondo, J. (2018). Do you Want to be a Cyborg? The Moderating Effect of Ethics on Neural Implant Acceptance. *Computers in Human Behavior*, *85*, 43–53. https://doi.org/10.1016/j.chb.2018.03.032

Reinares-Lara, E., Olarte-Pascual, C., Pelegrin-borondo, J., & Pino, G. (2016). Nanoimplants that Enhance Human Capabilities: A Cognitive-Affective Approach to Assess Individuals' Acceptance of this Controversial Technology. *Psychology & Marketing*, *33*(9), 704–712. https://doi.org/10.1002/mar.20911

Reinartz, W., Haenlein, M., & Henseler, J. (2009). An Empirical Comparison of the Efficacy of Covariance-Based and Variance-Based SEM. *International Journal of Research in Marketing*, *26*(4), 332–344. https://doi.org/10.1016/j.ijresmar.2009.08.001

Riek, L. D., & Robinson, P. (2008). Real-Time Empathy: Facial Mimicry on a Robot. *Workshop on Affective Interaction in Natural Environments (AFFINE) at the International ACM Conference on Multimodal Interfaces*, 1–5.

Ruiz-Mafe, C., Chatzipanagiotou, K., & Curras-Perez, R. (2018). The role of emotions and conflicting online reviews on consumers' purchase intentions. *Journal of Business Research*, *89*(January), 336–344. https://doi.org/10.1016/j.jbusres.2018.01.027

Russell, J. A. (1979). Affective Space is Bipolar. *Journal of Personality and Social Psychology*, *37*(3), 345–356. https://doi.org/10.1037/0022-3514.37.3.345

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Russell, J. A. (2003). Core Affect and the Psychological Construction of Emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145

Russell, J. A., & Mehrabian, A. (1977). Evidence for a Three-Factor Theory of Emotions. *Journal of Research in Personality*, *11*(3), 273–294. https://doi.org/10.1016/0092-6566(77)90037-X

Satterfield, T., Kandlikar, M., Beaudrie, C. E. H., Conti, J., & Herr Harthorn, B. (2009). Anticipating the Perceived Risk of Nanotechnologies. *Nature Nanotechnology*, *4*, 752–758. https://doi.org/10.1038/nnano.2009.265

Schicktanz, S., Amelung, T., & Rieger, J. W. (2015). Qualitative assessment of patients' attitudes and expectations toward BCIs and implications for future technology development. *Frontiers in Systems Neuroscience*, *9:64*. https://doi.org/10.3389/fnsys.2015.00064

Schifter, D. E., & Ajzen, I. (1985). Intention, Perceived Control, and Weight Loss: An Application of the Theory of Planned Behavior. *Journal of Personality and Social Psychology*, *49*(3), 843–851. https://doi.org/10.1037/0022-3514.49.3.843

Shim, S., Eastlick, M. A., Lotz, S. L., & Warrington, P. (2001). An online prepurchase intentions model: The role of intention to search: Best Overall Paper Award—The Sixth Triennial AMS/ACRA Retailing Conference, 2000☆11☆ Decision made by a panel of Journal of Retailing editorial board members. *Journal of Retailing*, *77*(3), 397–416. https://doi.org/10.1016/S0022-4359(01)00051-3

Stock, R. M., & Hoyer, W. D. (2005). An attitude-Behavior Model of Salespeople's Customer Orientation. *Journal of the Academy of Marketing Science*, *33*(4), 536–552. https://doi.org/10.1177/0092070305276368

Talukder, M. S., Chiong, R., Bao, Y., & Hayat Malik, B. (2019). Acceptance and Use Predictors of Fitness Wearable Technology and Intention to Recommend: An Empirical Study. *Industrial Management and Data Systems*, *119*(1), 170–188. https://doi.org/10.1108/IMDS-01-2018-0009

van der Heijden. (2004). User Acceptance of Hedonic Information Systems. *MIS Quarterly*, *28*(4), 695. https://doi.org/10.2307/25148660

Venkatesh, V. (2000). Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model. *Information System Research*, *11*(4), 342–365. https://doi.org/10.1287/isre.11.4.342.11872

Venkatesh, V., & Bala, H. (2008). Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences*, *39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, V., & Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, *46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer Acceptance and Use of Information Technology: Extending the Unified Theory od Acceptance and Use of Technology. *MIS Quarterly*, *36*(1), 157–178. https://doi.org/10.2307/41410412

West, J. (2016). Ability and Abnormality. In *(Master's Thesis)*. University of North Florida.

Wu, Y.-C., Tsai, C.-S., Hsiung, H.-W., & Chen, K.-Y. (2015). Linkage Between Frontline Employee Service Competence Scale and Customer Perceptions of Service Quality. *Journal of Services Marketing*, *29*(3), 224–234. https://doi.org/10.1108/JSM-02-2014-0058

Yahanda, A. T., Lafaro, K. J., Spolverato, G., & Pawlik, T. M. (2016). A Systematic Review of the Factors that Patients Use to Choose their Surgeon. *World Journal of Surgery*, *40*(1), 45–55. https://doi.org/10.1007/s00268-015-3246-7

Yang, H., Yu, J., Zo, H., & Choi, M. (2016). User Acceptance of Wearable Devices: An Extended Perspective of Perceived Value. *Telematics and Informatics*, *33*(2), 256–269. https://doi.org/10.1016/j.tele.2015.08.007

Yang, Q., Pang, C., Liu, L., Yen, D. C., & Michael Tarn, J. (2015). Exploring consumer perceived risk and trust for online payments: An empirical study in China's younger generation. *Computers in Human Behavior*, *50*, 9–24. https://doi.org/10.1016/j.chb.2015.03.058

Zuniga, J., Katsavelis, D., Peck, J., Stollberg, J., Petrykowski, M., Carson, A., & Fernandez, C. (2015). Cyborg beast: A low-cost 3d-printed prosthetic hand for children with upper-limb differences. *BMC Research Notes*, *8*(10), 1–8. https://doi.org/10.1186/s13104-015-0971-9

EXPLORING CO-DESIGN CONSIDERATIONS FOR EMBEDDING PRIVACY IN HOLOCHAIN APPS: A VALUE SENSITIVE DESIGN PERSPECTIVE

Paul d'Aoust, Oliver Burmeister, Anisha Fernando, Anwaar Ulhaq, Kirsten Wahlstrom

Holochain (Canada), Charles Sturt University (Australia), University of South Australia (Australia), Charles Sturt University (Australia), University of South Australia (Australia)

oburmeister@csu.edu.au; paul.daoust@holo.host; Anisha.Fernando@unisa.edu.au; aulhaq@csu.edu.au; Kirsten.Wahlstrom@unisa.edu.au

ABSTRACT

Supporting privacy in contexts mediated by technology is a persistent challenge and this paper sets out to advance practice through value sensitive design and a data sharing platform called Holochain. Values shape the way people behave and interact and support the formation of identity. Value sensitive design supports designers, developers and others in attending to values when creating new technologies. One such value is privacy which in this paper is seen as intrinsic to (and shaping of) social contexts, and ubiquitous in varying degrees. The paper supports developers and others engaged in the co-design of Holochain applications by describing which features of Holochain can be used to support privacy, offering six technical features for doing so.

INTRODUCTION

When introducing their book on value sensitive design, Friedman and Hendry (2019) quoted Winograd and Flores, "in designing tools, we are designing ways of being" (Friedman & Hendry, 2019, p1). Value sensitive design supports developers, designers and others in attending to human values in the design of technologies (Friedman & Hendry, 2012). Values are important in the design of technology because values shape the way people interact and behave, as well as supporting us in shaping our identities (Alshehri, Kirkham, & Olivier, 2020). However, values do not exist in insolation and may be in tension, for example, if developers are working on short deadlines, the provision of accessibility may be in tension with expedience. The value attended to in this paper is privacy.

Privacy has been discussed for centuries. Aristotle differentiated the *oikos* from the *polis* and considered the head of a household responsible for representing the views of the private household in the public sphere (as described in Molitorisz, 2020). Later, John Stuart Mill wrote, "That there is, or ought to be, some space in human existence thus entrenched around, and sacred from authoritative intrusion, no one who professes the smallest regard to human freedom or dignity will call in question," (Mill, 1848, p938). Later still, Warren and Brandeis suggested that privacy was a right worthy of protection under the law (Warren & Brandeis, 1890). In 1929, Ragland perceived the challenges to privacy suggested by technologies and made the prescient suggestion that privacy may be of unusual importance in years to come, given the technological advances of his time (Ragland, 1929).

More recently, privacy has been variously described as the control of data in the pursuit of self-determination (Westin, 1967), restricting access to self and data (Moor, 1990), the capacity to trade (or otherwise) privacy and data as commodities (Posner, 1977), a social good unique to specific contexts (Burmeister, Islam, Dayhew, & Crichton, 2015; Nissenbaum, 2009), and as being shaped by the technologies forming the infosphere (Floridi, 2005). Today, privacy is understood to arise in diverse

social contexts (Nissenbaum, 2009), taking diverse and pliant forms appropriate to those contexts (Wahlstrom, Fairweather, & Ashman, 2017) and is often mistakenly conflated with security and confidentiality (Mittelstadt, Fairweather, McBride, & Shaw, 2013; Wahlstrom & Fairweather, 2013). Regardless of whether technology mediates a social context, privacy is intrinsic to that context and pliant. Privacy is both shaped by social contexts and shaping of it (Wahlstrom et al., 2017), with some social contexts calling for no privacy while others call for much privacy. Thus, privacy is ubiquitous but in varying degrees, and frequently in tension with values such as seeking to serve shareholders by deriving market value from data assets.

In some cases, information technologies provide high levels of data privacy and in others, very little privacy is available. Regardless of how much or little privacy is afforded, pliancy in data privacy is rarely available. The right to be forgotten (Cellan-Jones, 2014) affords a type of pliancy in data privacy because under this right, someone may request that data be removed from search indexes and thus increase their data privacy. Wahlstrom, Ulhaq, and Burmeister (2020) considered the right to be forgotten in the context of an emerging peer to peer platform called *Holochain*. This paper extends that work by exploring considerations relevant to enabling privacy in Holochain applications (called *hApps*) through values sensitive co-design.

The next section provides an overview of data architectures followed by a description of Holochain, focussing on agency, privacy, and anonymity. This is followed by a discussion of how privacy can be supported with Holochain, which suggests six privacy design considerations for developers and eight community co-design considerations. This is followed by a section that discusses the six privacy design considerations in detail, prior to concluding the paper.

HOLOCHAIN

Applications that use a client/server architecture preserve data integrity by erecting a fortified wall around the data, consist of cybersecurity measures, data access policies, and validation and access control rules encoded in the application itself. However, the details of these protection measures are opaque, which may be unsatisfying for those seeking data privacy through agency over how the data describing them is used. Furthermore, this enclosure of data creates asymmetries in access to information, which in turn foster asymmetries in power (Janssen, Cobbe, & Singh, 2020).

With the introduction of two fundamental cryptographic tools: secure hashing algorithms and public-key encryption, distributed computing systems have become increasingly viable. These have offered solutions to key problems, such as verifiable, tamper-proof data for sharing state across distributed system nodes and data provenance validation using digital signature algorithms.

These systems can be divided into two types: data-centric and agent-centric. Data-centric systems aim to create a single shared data reality for all nodes (blockchains are examples of data-centric systems). Agent-centric distributed systems are concerned with allowing nodes to share constantly changing data (Zaman, Khandaker, Khan, Tariq, & Wong, 2021).

Blockchain is a transparent alternative to client/server applications, allowing equal access to both the data it stores and the details of the algorithms by which data integrity is preserved. However, this is possible only because all data is contained in a global public ledger that is replicated, inspected, and audited by all parties with a copy of the software that encodes the data integrity algorithms. Fine-grained access control to specific portions of the ledger is by definition impossible.

Holochain is an agent-centric distributed generalised computing system in which nodes can still participate in the system as a whole while not needing to maintain the same chain state as the other

nodes (Harris-Braun, Luck, & Brock, 2018). A Holochain application (a *hApp*) "… consists of a network of agents maintaining a unique source chain of their transactions, paired with a shared space implemented as a validating, monotonic, sharded, distributed hash table (DHT) where every node enforces validation rules on that data in the DHT as well as providing provenance of data from the source chains where it originated" (Harris-Braun et al., 2018, p4). In Holochain's DHT and its proof by enforcement. all elements of the DHT can only be modified monotonically; that is, elements can only be added to DHT and not removed.

Holochain and agency

Holochain starts with a fundamentally different perspective from either blockchain or client/server architectures: an acknowledgement that all information has its origin in the subjective experience of the agent (usually a person) producing it, and the argument that data separated from its provenance has lost a critical part of its meaning (Harris-Braun, 2021). Holochain is designed around this perspective.

In putting agency rather than data at the centre of Holochain, there was no intention to explicitly enhance individual privacy or freedom, however these features are natural side-effects of this orientation (Harris-Braun, 2021).

Each participant in the system exercises their agency via a computer device under their control, recording their actions as entries on a personal digital journal. These entries are cryptographically signed by a private key, which provides the aforementioned provenance context. When the application requires them to share these entries publicly, they do so by broadcasting to a subset of their peers who use the same application. Those peers then take responsibility for validating, storing, and serving the entries to others.

Holochain and privacy

There is a dynamic tension between individual and collective agency, which application developers and other stakeholders must consider in their designs. This becomes especially apparent with respect to privacy. When an individual keeps their information secret, it is naturally as private as possible. However, a networked application is useful precisely because it allows information to be shared. This requires the individual to divulge some of their information in order to derive utility from the application.

When someone discloses their information to others, whether peers in a peer-to-peer digital network, to a centralised server, or to other people, it is witnessed and becomes in one sense part of a collective memory. It may never be completely forgotten; even if it is diligently removed from all digital databases, backups, and caches, it may still live on in the memories of those who saw it.

The medium in which a piece of information is preserved imparts an amount of 'friction' (Floridi, 2005) or 'greasing' (Moor, 1990) to the movement of that information, regardless of whether the medium is a technological context. When someone is asked to forget information shared with them, the intention may not be to request they remove it from their memory, but to refrain from circulating it in low-friction media; for example, to refrain from speaking it to others. This results in enhanced privacy because the movement of the information is impeded.

There may also be cases where 'my information' cannot be distinguished from 'our information' without damaging the integrity of dependent information and thus causing material harm. An example

is a 'mutual credit' or 'barter exchange' network, in which each member keeps their own ledger of debits and credits. Often, ledgers are visible to all other members, in order to facilitate mutual auditing and credit checking.

Consider a scenario in which Alice transfers some of her credit to Bob in exchange for goods, which gives Bob a positive balance that he can then spend with others. In a Holochain-based implementation, each transaction involves Alice and Bob mutually creating and countersigning a transaction record, which is then written to their respective ledgers. The transaction record, as a product of their interaction, belongs to both of them. If Alice were to leave the system and ask Bob to 'forget' her transaction with him, Bob's ledger would no longer appear to have sufficient positive balance to spend.

In light of these privacy issues in a peer-to-peer network, it can be suggested that the most optimal privacy-preserving designs will be highly tailored to the social context of the group for whom the application is developed. This coheres with the intent of a Holochain application: if the ostensible aim of the executable code is to embody the norms of a group of users, then strong feedback loops should exist between those users and the developers writing the code, ideally using a participatory design process such as the Value Sensitive Action-Reflection Model described below.

Holochain and anonymity

On the other hand, Holochain does not easily support privacy through anonymity, as it was designed to support accountability among individuals who consent to participate in a context. An agent ID (the public key of an asymmetric key pair) is a long-lived identifier. Although the Holochain protocol does not prevent users from creating an arbitrary number of anonymous IDs, an application that permits this may be more vulnerable to an attacker running multiple nodes in an attempt to overwhelm the network's ability to assure data integrity; a Sybil attack (Douceur, 2002).

A Holochain application consists of a set of rules that define valid actions an agent can take. This is not unique; almost all software that deals with user input incorporates validation rules. What makes Holochain and other peer-to-peer protocols unique, especially compared to client/server applications, is that every participant has a copy of these rules on their own device. This allows them to ensure that their own actions are valid and to check the validity of others' actions. What emerges from these two properties is that the individual agents using the application form a cohesive group, defined by their mutual consent to be bound by the application's validation rules. These rules become a digital, executable encoding of the group's norms (Lessig, 2000). For this reason, the Holochain protocol is described as *social DNA* and the core executable code of a Holochain application is called a DNA bundle.

SUPPORTING PRIVACY WITH HOLOCHAIN

hApps have the potential to support privacy because Holochain provides scope for the appropriate consideration of social contexts. As noted in the introduction, social contexts give rise to privacy, regardless of whether a context is mediated by technology. Key privacy by design considerations in hApps can be embedded by creating and configuring contexts that uphold the privacy expectations of social contexts, such as the right to be forgotten. From a developer's perspective, there are six design considerations, namely:

Entry visibility Defining data types as private, public, or public but encrypted in data schemes.

Membership membranes Participants' access to a hApp's network space can be specified in code given that all public data is visible in this space.

Partitioned data Use of two sharding methods: each party maintains their own journal, and each separate entry and header in a journal can be viewed, validated and stored by a small sample of other participants.

Capability-based security Access to private entries on a journal can be approved via use of capability tokens that may be non-transferable.

API membrane Only public data is verified or approved by the network, with the definition of valid data being determined by the developer and the participants' contexts.

Withdraw and purge Two new techniques to remove data are theorised with the implementation details pending: withdraw (redact data formerly published) and purge (offer a participant to mark another's data for deletion).

However, while developer requirements are important, the requirement to consider design needs from the community of people running or accessing hApps supports the balancing of data integrity with privacy. This ethical tension can be addressed by applying the value sensitive action-reflection model, a value sensitive design tool (Friedman & Hendry, 2019), see Figure 1. We suggest using the value sensitive action-reflection model recommended by Yoo et al. (2013) for the co-design of hApps in order to address the privacy needs arising in social contexts. The development of designer and stakeholder prompts is noted as a future research opportunity.

For example, the design needs of the people running or potentially accessing hApps should be considered. One approach to raise these design considerations draws on the core ICT values of privacy suggested by Huldtgren (2014): security, ownership, universal usability, autonomy, trust, accountability and human welfare. Hence, when developing a co-design space for hApps, these community-based design considerations could be explored:

Privacy This value is of core focus to this social context, and privacy-by-design criteria are embedded as listed in the developer considerations above. hApps have the potential to appropriately apply contextual privacy specific to social contexts. To understand the nuances of these social contexts when designing and to increase uptake across practical applications, participants may need data ethics literacy skills. Efforts to grow the awareness and implementation of privacy-related data knowledge and skills are needed to maintain the integrity of hApps.

Security What are potential security risks that need to be considered from a hApp community practitioner's perspective? For example, if social engineered cyber-attacks have the potential to input unauthorised data into the holochain, some privacy-by-design safeguards need to be considered when codesigning in practice.

Figure 1. Evolving the Co-Design Space for Holochain, adapted from Yoo, Huldtgren, Woelfer, Hendry, and Friedman (2013).

Ownership As a disruptive approach to managing and storing data, how is this decentralised form of data ownership perceived by potential community users? For example, community-based hApps may need to be designed with common public-access protocols in place both from technical and human perspectives.

Universal Usability Can anyone use a hApp or are specific types of knowledge and skills required? How can these competencies be inclusively acquired? For example, if the hApp is designed for a context, considerations for accessibility and inclusiveness may need to be incorporated in subsequent iterations of the hApp.

Autonomy How can hApp users practice autonomy over managing data choices? For example, users may need to think of and use the data embedded as having collective privacy (i.e. “our data”) as opposed to individual privacy (i.e. “my data”).

Trust What ensures the trustworthiness of hApp and the integrity of data managed in these transactions? For example, end-users, designers and decision-makers will need to adapt to new practices and measures of trustworthiness and integrity pertaining to the data held in these hApps.

Accountability and Responsibility Which stakeholder(s) are accountable and responsible for the integrity of these hApps? What are mechanisms to practice these values? For example, organisations developing hApps may need to develop capability by having digital or data ethics designers incorporated in their project development teams to embed privacy-by-design principles.

Human Welfare Are there any potential harms propagated or introduced by managing data in hApps? For example, the value sensitive action reflection model is useful to apply in this context to reflect on potential instances where the contextual nature of privacy is upheld, disrupted or strained such as via a value tension. The consideration of these central ICT values is not an exhaustive list but is recommended as a useful place to start in unpacking the ethical tensions in emerging technologies such as hApps. The potential of these emerging technologies to appropriately address the need for embedding social contexts for privacy and further the contextual relational nature of privacy should be explored.

SIX PRIVACY DESIGN CONSIDERATIONS

To support the value sensitive design of hApps, co-designers should have a strong grasp of the basic building blocks of Holochain and how they affect privacy, as well as the tensions between privacy and data integrity mentioned above. This knowledge ought to serve as foundations for useful design prompts to aid in the participatory design process. This section details the six design considerations that affect privacy that were identified above.

Entry visibility

An individual's journal consists of entries linked together cryptographically by headers. For each entry (not including system entries), a data type or schema can be defined, along with the type's *visibility*. The visibility attribute defines whether entries of this type are shared with a random subset of peers who are using the same application. These peers are called *authorities,* by virtue of having been selected to validate, store, and make the data available to any peer who has a reference to it. This mechanism builds a distributed hash table (DHT) of shared data among the users of the application. The two visibility options for an entry type are:

> *Private*, in which only the header is shared, while the entry content is kept locally on the individual's device; or
>
> *Public*, in which both the entry header and content are shared.

The distinction between these two can be compared to a card game in which some cards are dealt face-up and some are dealt face-down for one player to put in their hand. All players can see every card that has been dealt; not all players can see the contents of all cards (Brock, 2009).

The availability of this option stands in contrast to most public blockchains, which typically require all data to be made public for validation by all peers. While metadata can be written to blockchain transactions to refer to the existence of off-chain data, similar to private entries in Holochain, this is an ad-hoc solution that requires integration with another technology in order to store said data.

There are two compromises that must be made in deciding between private and public visibility; this is noted as possible design guideline. First, private data suffers from reduced availability, as it requires the author to be online if any peer needs to request it. Second, it cannot be publicly validated, so it cannot contribute to the network's aggregation of metadata that indicates what and who can be trusted.

As previously mentioned, as a third option public data can also be encrypted before sharing in order to 'hide it in plain sight'. Available to blockchains as well, this option allows data to be accessed by others who know how to decrypt it, even when the author is offline. Traditional symmetric-key encryption can be used, although this once again prevents authorities from being able to validate it.

Novel encryption schemes, such as zero-knowledge proofs (eg ZK-SNARKS, pioneered with the ZCash blockchain), allow data to remain secret while still being subject to validation; this is also an option for Holochain applications. However, such encryption schemes may incur high computational overhead and not all validation problems can be modelled with zero-knowledge proofs.

Membership membranes

Holochain anticipates the need to define appropriate privacy norms for different contexts. As each DNA bundle and its network of users represent a context, it is possible to define a 'membrane' that determines who does and does not belong to the context. This is implemented via a validation function on an entry near the beginning of a journal, the 'joining proof'. This proof can contain credentials such as an invite code, an attestation signature from an existing member, or a proof of identity or membership in the corresponding human network.

Context membership can be further managed by ejecting an individual from the context. This may be required as a response to the intentional production of invalid data, which would indicate an individual has tampered with their copy of the software. It can also be used in non-adversarial situations, such as an employee leaving a company and having their access to company data revoked.

Partitioned data

Data in Holochain is partitioned in three ways. Firstly, as noted above, the individual's journal gives them the power to create their own data and permits them to keep some of that data unpublished.

Secondly, each application has its own network with its own store of public data (DHT). Here, 'public' indicates accessibility to other authenticated context members within the application's membrane, not accessibility to the world. Networks are isolated from each other and distinguished by the cryptographic hash of their DNA bundles' executable code. This can also be exploited to create multiple separate contexts operating with the same executable code, by changing an insignificant detail such as the application's title or ID that causes the hash to change.

Thirdly, the authorities selected to validate, store, and serve each piece of public data are only a subset of the entire context, and are selected randomly. Any and likely all agents will be selected at various times to be authorities for at least some pieces of public data. This distribution, called sharding, scatters data throughout the network. This reduces the amount of data that any agent is able to easily analyse. Additionally, a piece of data is referenced via an opaque ID (the cryptographic hash of the content) and cannot be retrieved by an agent unless they know this ID.

These are weak privacy measures, however, as any agent can enlist themselves as an authority for an arbitrarily large portion of the public data set. It does, however, place a greater computation and storage burden on such a peer, which could potentially serve as a deterrent to anyone who wishes to capture and analyse the entire data set.

Capability-based security

An individual can generate *capability grants* that give selective permission for peers to call specified public functions in their running DNA instance. In a sense, the individual is delegating some of their agency to others, as these functions have all the privileges the owner of the application instance enjoys. The developer can use this feature to, for example, create 'getter' (accessor) functions that return private journal entries.

A capability can be granted with one of three levels of access restriction:

- An *unrestricted* grant allows anyone to call the function for which the capability is being granted and is often applied to functions that allow peers to negotiate the granting of capabilities with tighter access restriction.

- A *transferable* grant is coupled with a secret token, and only permits peers who possess the token to call the function.
- An *assigned* grant is coupled with both a secret token and a list of peer IDs, and only permits access for peers who possess the token and whose ID appears in the list.

These grants can be created and revoked at any time, according to the functionality that the developer has designed into the application. This allows much finer-grained access control than the simple public-versus-private visibility described in point 1, although data protected and accessed in this manner is private and thus subject to the compromises described in that section.

API membrane

An application's DNA bundle consists primarily of functions that abstract over the low-level Holochain functionality and present a coherent view of the underlying data. In this sense, such functions can be considered a data access layer or API. Their chief purpose is to allow the owner of a running application instance to exercise their agency via a user interface (although, as point 4 mentions, capability tokens can be used to delegate that agency to others).

It must be noted that, when data is published to the DHT, it cannot properly be deleted or modified. This is due to the fact that the DHT is a monotonically increasing set that merely aggregates data points; this makes it easier to replicate data (Hellerstein & Alvaro, 2019) and maintains the historical record necessary to ensure distributed data integrity. However, data can be marked as *deleted* or *updated* and this appears as a piece of metadata attached to the entry that marks it as obsolete and, in the case of *updated*, points to a replacement entry.

When accessing public DHT data, the DNA bundle's functions can manipulate the result set before returning it to the UI. Specifically, they could be written to honour the deleted or updated status of entries and filter them from the return value. While this does not guarantee that the data is inaccessible, it does require a considerable amount of technical skill to bypass the Holochain runtime and access the raw data. In light of the view that privacy is enhanced and eroded in part by friction/greasing, it may not be absolutely necessary to completely eliminate the data in question; the aforementioned could potentially satisfy, for example, the right to be forgotten in an acceptable fashion for some applications.

Withdraw and purge

The Holochain protocol implements two new operations for the shared database: *withdraw*, which allows an author to ask validators to remove data they mistakenly published; and *purge*, which allows anyone to mark anyone else's data for removal. These are intended to be 'true' deletion operations for public DHT data, causing all compliant peers to honour the requests by erasing their copies of the data from their devices. While these operations are often used for the removal of errors or illegal content, we anticipate that they could also be used to exercise contextual privacy norms in a way that aligns more closely with a user's natural expectations. It should be noted, however, that in a peer-to-peer system these operations merely constitute a polite request and cannot be algorithmically forced upon non-compliant peers.

We also note that such capacities may be incompatible with the needs of a given context. As seen previously in the example of a mutual credit network, data that becomes part of the public record, i.e.,

'our information', can become highly interrelated. In these cases, removing one entry might cause entries that depend on it to become invalid. Additionally, even for data that exists independently of other data, the context's rules may preclude that possibility, such as for a federal voting application that requires immutable records in order to confirm the results of an election.

Finally, a meta-consideration involves the upgrading of context rules. Effective privacy adapts to changing social contexts, so any privacy rules embedded in a DNA bundle also need to adapt if they are to serve evolving or emerging privacy norms. As a peer-to-peer application only exists when there are at least two individuals actively running instances of it, those individuals can agree to upgrade to a new application with a new set of rules, split into multiple groups, or elect to stay with the old rules if they prefer them. This can be assisted through upgrade routines in a DNA bundle, which may require that sufficiently flexible group governance processes for upgrades be designed into the first iteration of the application. This may offer a more flexible option than public blockchain platforms which offer an unchangeable set of base rules and do not allow application-specific rules (smart contracts) to be modified, although we also note that blockchain developers are beginning to establish design patterns that allow smart contracts to be retired or superseded by new versions.

CONCLUSION

Values are integral to the way people interact and behave. Value sensitive design supports developers and others in attending to values as they undertake the co-design of technologies. With respect to hApps, it is important to focus attention on privacy, security, ownership, universal usability, autonomy, trust, accountability and responsibility, and human welfare.

The support of privacy in contexts mediated by technology has been a persistent problem, with regulatory approaches failing to support contextual nuances and evolving privacy norms, and technologies failing to offer sufficient diversity and flexibility in privacy options. Unlike Blockchain which does not support privacy effectively, Holochain is a data sharing technology offering six features that may be leveraged by developers to create applications that support nuanced and contextual privacy norms while sharing data in non-repudiable ways. A future research opportunity is the investigation of the extent to which these features can be successfully leveraged in value sensitive design co-design projects.

KEYWORDS: value-sensitive design (VSD), privacy, Holochain.

REFERENCES

Alshehri, T., Kirkham, R., & Olivier, P. (2020). *Scenario Co-Creation Cards: A Culturally Sensitive Tool for Eliciting Values.* Paper presented at the Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.

Brock, A. (2009). Differences Between "Open Source" and "Open Currency". Retrieved from https://www.artbrock.com/2009/05/12/differences-between-open-source-and-open-currency#digging-deeper-into-open-data

Burmeister, O. K., Islam, M. Z., Dayhew, M., & Crichton, M. (2015). Enhancing client welfare through better communication of private mental health data between rural service providers. *Australasian Journal of Information Systems, 19*. https://doi.org/10.3127/ajis.v19i0.1206

Cellan-Jones, R. (2014). EU court backs 'right to be forgotten'in Google case. *BBC News (14 May 2014) online: BBC News Europe* http://www.bbc.com/news/world-europe-27388289

Douceur, J. R. (2002). *The sybil attack.* Paper presented at the International workshop on peer-to-peer systems.

Floridi, L. (2005). The Ontological Interpretation of Informational Privacy. *Ethics and Information Technology, 7*(4), 185-200. https://doi.org/10.1007/s10676-006-0001-7

Friedman, B., & Hendry, D. (2012). *The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations.* Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Friedman, B., & Hendry, D. (2019). *Value sensitive design : shaping technology with moral imagination*. Cambridge: MIT Press.

Harris-Braun, E. (2021). Decentralized Next-level Collaboration Apps with Syn. Retrieved from https://blog.holochain.org/decentralized-next-level-collaboration-apps-with-syn/#back-story-agent-centric-data-enables-collaboration

Harris-Braun, E., Luck, N., & Brock, A. (2018). Holochain: scalable agent-centric distributed computing. Retrieved from https://github.com/holochain/holochain-proto/blob/whitepaper/holochain.pdf

Hellerstein, J. M., & Alvaro, P. (2019). Keeping CALM: when distributed consistency is easy. *arXiv preprint arXiv:1901.01930*.

Huldtgren, A. (2014). Design for Values in ICT. In M. J. Van Den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook on ethics, values and technological design: sources, theory, values and application domains* (pp. 1-24). Netherlands: Springer.

Janssen, H., Cobbe, J., & Singh, J. (2020). Personal information management systems: A user-centric privacy utopia? *Internet Policy Review, 9*(4), 1-25.

Lessig, L. (2000). Code is law. *Harvard magazine, 1*(2000).

Mill, J. S. (1848). Principles of political economy with some of their applications. *Social Philosophy*.

Mittelstadt, B., Fairweather, B., McBride, N., & Shaw, M. (2013). *Privacy, risk and personal health monitoring.* Paper presented at the ETHICOMP, Kolding, Denmark.

Molitorisz, S. (2020). *Net Privacy: How we can be free in an age of surveillance*: NewSouth.

Moor, J. (1990). The ethics of privacy protection. *Library Trends, 39*(1), 69-82.

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, US: Stanford Law Books.

Posner, R. (1977). The right of privacy. *Ga. L. Rev., 12*, 393.

Ragland, G. (1929). The Right of Privacy. *Kentucky Law Journal, 17*, 85-122. doi:citeulike-article-id:13666094

Wahlstrom, K., & Fairweather, N. B. (2013). *Privacy, the Theory of Communicative Action and Technology*. Paper presented at the ETHICOMP, Kolding, Denmark.

Wahlstrom, K., Fairweather, N. B., & Ashman, H. (2017). *Brain-Computer Interfaces and Privacy: Method and interim findings*. Paper presented at the ETHICOMP, Turin, Italy.

Wahlstrom, K., Ulhaq, A., & Burmeister, O. (2020). Privacy by design. *Australasian Journal of Information Systems, 24*.

Warren, S., & Brandeis, L. (1890). The Right to Privacy. *Harvard Law Review, 4*(5), 193-220. doi:citeulike-article-id:6597501

Westin, A. (1967). *Privacy and Freedom*: Atheneum.

Yoo, D., Huldtgren, A., Woelfer, J. P., Hendry, D. G., & Friedman, B. (2013). *A value sensitive action-reflection model: evolving a co-design space with stakeholder and designer prompts*. Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Zaman, S., Khandaker, M. R., Khan, R. T., Tariq, F., & Wong, K.-K. (2021). Thinking Out of the Blocks: Holochain for Distributed Security in IoT Healthcare. *arXiv preprint arXiv:2103.01322*.

ETHICAL RESPONSIBILITY IN SPACE EXPLORATION

Zachary J. Goldberg

Trilateral Research (Ireland)

zachary.goldberg@trilateralresearch.com

INTRODUCTION

The ethical questions arising in the context of current and future space travel and exploration are as abundant as they are complex (Schwartz and Milligan 2016). This is owing to both the increasing number of technological possibilities opening opportunities for human exploration and exploitation of space, other planets and asteroids in addition to the myriad of relevant scientific and ethical factors that remain unknown about these activities. Much of the literature focusing on an ethics of space exploration has attempted to apply and analyse the traditional (Western) ethical theories of deontology, consequentialism and virtue theory to the context of space travel and exploration, and then discuss which of these theories would be the most appropriate ethical guide for human activities in space (e.g. McArthur and Boran 2004, Arnauld 2011, Green 2014). In this paper I argue that these discussions are invaluable, not because they provide a concrete ethical guide to follow, *but because they do not*. Understanding ethical theories is only the first step in cultivating an ethic of personal responsibility that focuses on: a) questioning the nature of ethical values, and, b) questioning one's abilities to act ethically especially in the context of significant unknowns. I conclude that cultivating personal responsibility as it is here described is critical for stakeholders involved in current and future human activities in space.

ETHICAL THEORIES AND SPACE EXPLORATION

As more money is invested in space, as more people—both professional astronauts as well as tourists—travel to space, as our spaceships are able to venture farther out into the galaxy, as we contemplate mining extra-terrestrial resources for use on earth, or consider terraforming and colonizing Mars or other celestial bodies, we are confronted with a myriad of ethical questions, both theoretical and practical:

Does the space environment (including the solar system and beyond) contain anything of inherent value? Do we have an ethical obligation to limit our activities on celestial objects such as asteroids, comets, moons, or planets? Or are they there for us to research and exploit? Are we ethically permitted to take resources from the moon or other planets for use on earth? Should we preserve pristine space environments? If we discover extraterrestrial life, including microbial, would it deserve our moral consideration? For what reasons? To what extent? If long duration space flight becomes technologically feasible, would it be justifiable to send humans into space for years or decades? What are the risks involved? If we discover that another planet, e.g. Mars, would be habitable if we drastically altered the landscape, also known as terraforming, are we justified in doing so? What challenges would space colonies face both in terms of physical survival but also in terms of psychological hardships? Is the current budget allocated for space exploration justifiable when there are injustices on earth that need urgent attention? How can we clean up the vast amount of space junk in orbit? How can we avoid future missions adding to this pollution? How can we further exploit satellites for earth observation to combat climate change, or to provide digital education to the world's

population? Does space belong to no one or to everyone? What are the legal ramifications of this answer regarding property and ownership? Is outer space a land-grab where prospectors can claim planetary resources on a first-come first-served basis? Which regulatory protocols do we need in place? How can we foster continued international collaboration in space? Is exploration a good in its own right, or is it only justifiable if it yields actionable research? Are we justified in asking astronauts to engage in such high-risk activity which literally changes their bodies in terms of fluid distribution, loss of body mass, and sleeplessness?

The ethical, scientific, legal and societal significance of these questions remains indisputable. Unsurprisingly, scholars in the domains of space ethics particularly, as well as the ethics of technology more generally, have approached these questions and others about the impact of new technologies by appealing to influential ethical theories (Green 2014; Schwartz and Milligan 2016). This approach is not surprising because consequentialism and utilitarianism, Kantian deontology, and virtue ethics are the most influential ethical theories in the Western tradition precisely because they are insightful, have continually resolved various ethical quandaries, can each appeal to common intuitions about ethical action and decision making, identify values important to human interaction with other humans, with animals and with the environment, and have been tried and tested by philosophers for centuries. To these principal theories, we can also add more recent developments in philosophical and applied ethics that also provide insight to the issues and questions concerning space travel and exploration, namely, the ethics of care, principlism, multiple theories of justice, and differing approaches to issues in environmental ethics (Daly and Frodeman 2008).

Scholarship engaging these theories in the context of human activities in space has provided an essential first step in assessing the ethical issues of space travel and exploration. For example, adopting a consequentialist approach equips us with the conceptual tools necessary to assess the consequences of space activities such as long-duration human space flight, colonizing and terraforming other planets, or weighing the costs of space tourism and increasing amounts of space debris (Baum 2016). Although consequentialists have offered differing definitions of good and bad, e.g. pleasure/pain, preference dis/satisfaction, ecosystem flourishing/failing, they are united in arguing that human actors are obligated to maximize good consequences and minimize bad ones. As a result, though we might not always know with certainty what the consequences of a certain space activity will be, consequentialism provides us with a clear principle to guide our actions and decision making in space.

Likewise, Kantian deontology can furnish an explicit ethical rule to steer human space activities. Kant's categorical imperative instructs human actors both to treat with respect other beings with the capacity for practical reasoning as well as to universalize ethical decisions by avoiding making oneself or one's actions an exception to an ethical rule (i.e., the moral law). In the space context, treating other autonomous, rational beings with respect is relevant for considering the health and mental wellbeing of astronauts travelling to and through outer space or when contemplating a long-duration flight or stay in space. According to this framework, space policy and decision makers ought to never treat astronauts as a mere means to their own goals even if doing so would yield positive consequences. Juxtaposed with Elon Musk's recent statement that "a bunch of people will probably die" in the first crewed missions to Mars, the Kantian framework reveals its relevance (Gohd 2021). Additionally, under the Kantian approach, companies planning to take tourists into space are obligated to treat their customers with dignity and respect, never trading off these values for profit or other reasons. In practical terms, acknowledging these values and their inviolability might help companies shift focus from profitability to protecting individuals' autonomy by drafting robust information and consent sheets to allow potential customers to make truly informed and autonomous decisions before

participating in space activities.[15] Finally, although it is less likely to occur, at least in the near future, the categorical imperative could guide our actions should humans encounter rational, autonomous alien beings.

Virtue ethics can also contribute to our ethical decision making in the context of human activities in space. In contrast to consequentialists or Kantian deontologists, virtue ethicists argue that good actions emerge from good characteristics, and, as a result, they identify which good characteristics one ought to cultivate. These good characteristics, or "virtues", are traits that reflect an excellence in practical wisdom. Aristotle names courage, honesty, and temperance among others as virtues that reflect such an excellence. He defines each virtue as a mean between two extremes, as it is extreme behaviour that causes unethical action. For example, courage is the mean between cowardice and recklessness, and temperance is the mean between self-indulgence and asceticism. The way to cultivate these virtues is through habit and life-long practice.

One might wonder whether an ethical theory first formulated in ancient Greece is relevant to evaluate human activities in space which are made possible by technologies unimaginable in Aristotle's era. Responding to this query, the philosopher Shannon Vallor argues that virtue ethics is superior to consequentialism and Kantian deontology in the technological age in large part due to the vast uncertainty that new technologies elicit. Virtue ethics helps individuals cultivate the ability to make ethically good decisions in the future even if the precise details of this future are currently unknown. Vallor contends that the virtues, such as those mentioned above, are reliable guides for ethical action and decision making regardless of context insofar as they constitute moral expertise.

Moral expertise thus entails a kind of knowledge extending well beyond a

> cognitive grasp of rules and principles to include emotional and social intelligence: keen awareness of the motivations, feelings, beliefs, and desires of others; a sensitivity to the morally salient features of particular situations; and a creative knack for devising appropriate practical responses to those situations, especially where they involve novel or dynamically unstable circumstances. (Vallor 2016, 26).

Novel or dynamically unstable circumstances describes many space activities such as terraforming and colonizing Mars. If Aristotle and Vallor are correct, by cultivating virtue, or moral expertise, individuals would be equipped to make ethically right and good decisions even in new and uncertain circumstances.

Despite the ethical insights that these theories contribute to our thinking about human activities in space, each theory faces critical objections and counterexamples. Consequentialism obligates us under some circumstances to sacrifice the few for the many, Kantian deontology rejects the possibility of trading off slight transgressions of a person's autonomy or dignity for significantly good consequences that could benefit humanity, and virtue ethics has been criticized as not being sufficiently action-guiding (McElreath 2017). Furthermore, each of these ethical theories is based in knowledge—knowledge of human nature, of what has intrinsic value, of the nature of the right or the good, or the identification of those character traits that result in good and right actions. Even though novelty and uncertainty of circumstances is acknowledged by Vallor, she maintains that moral expertise functions well in these circumstances. However, many of the foreseeable human activities in space are so radically different than their terrestrial counterparts that reliance on ethical knowledge to adjudicate ethical dilemmas and serve as a trusted guide in the context of space is worrisome at least.

[15] For an assessment of the legal implications of space tourism including information and consent, see Matignon 2019.

RESPONSIBILTY AS ETHICAL QUESTIONING, NOT ETHICAL EXPERTISE

Despite this worry, the criticisms directed toward these ethical theories ought not be seen in a negative light. Rather, understanding each of their positive aspects followed by reflection upon their individual inadequacies in the context of human activities in space, will prove essential to the cultivation of an ethic of responsibility for space exploration. Continued reflection will reveal that these ethical theories are not incorrect, but each is insufficient on its own to provide a comprehensive approach to the ethics of space. The subsequent awareness of the reasons for this insufficiency is, paradoxically, a significant ethical step; it is one that leads not to ethical expertise, but to ethical responsibility instantiated through the activity of inquiry. Given the significant unknowns of future human activities in space—will we encounter other life and in what form? will we go on long-duration space flights and what biological and psychological affects do such flights have? will we colonize other planets? will humans be born on other planets? how will tourism in space develop?—it is not moral expertise that will prove necessary, but a commitment to asking ethical questions.

There are two lessons from the history of philosophy that inspire this current position. Socrates communicates the first lesson through his method of interaction with his interlocutors. Famously, Socrates claimed that his awareness of ignorance, his own and others', made him the wisest person in Greece. This perspective is reflected throughout the *Dialogues* and continues to be central to Socrates' philosophical legacy (see, Drengson 1981). It is critical to note that Socrates' awareness of his own and others' ignorance is not a static position at which point he decides to rest comfortably simply knowing that he is ignorant. Rather, it is an awareness that is supported by continuous ethical questioning. As such, it is a process, or an activity. It is precisely this commitment to actively asking questions that has cemented Socrates as the "Father of Western Philosophy"—a tradition built upon inquiry and investigation. However, the activity of asking questions motivated by awareness of one's own and others' ignorance is not simply an epistemic position; it cultivates an ethic of responsibility.

To see this connection to responsibility, let us examine Socrates' interaction with Euthyphro. Socrates' examination of Euthyphro takes place just as Euthyphro has deposed murder charges against his own father. Murder was considered at the time to be a religious offense, an act of impiety, but a son taking such actions against his father could also be consider impious. As Socrates plainly points out, before one does something as drastic as reporting one's own father for the crime of impiety, one ought to know what impiety is.

> Socrates: What is your case, Euthyphro? Are you the defendant or the prosecutor?
> Euthyphro: The prosecutor.
> Soc: Whom do you prosecute?
> Euth: One whom I am thought crazy to prosecute.
> Soc: Are you pursuing someone who will easily escape you?
> Euth: Far from it, for he is quite old.
> Soc: Who is it?
> Euth: My father.
> Soc: My dear sir? Your own father?
> Euth: Certainly.
> Soc: What is the charge? What is the case about?
> Euth: Murder, Socrates.
> Soc: Good heavens! Certainly Euthyphro, most men would not know how they could do this and be right. It is not the part of anyone to do this, but of one who is far advanced in wisdom. (Plato 1997a, 3e-4b)

As is his custom, Socrates proceeds to listen to Euthyphro's definition of piety/impiety and expose its logical mistakes. Simply by questioning Euthyphro's ethical stance—his understanding of the value of piety and his confidence in his supposed ethical obligation to report his own father—Socrates reveals that Euthyphro's confidence in his ability to engage in an ethical assessment is sorely misplaced. The reader is left with the clear impression that Euthyphro's actions are irresponsible and unethical, but not because he is a malicious person. To the contrary, he wants to do the right thing. He endeavours to take ethical action despite not knowing what the ethically good action is in his circumstances, and more importantly, not knowing that he does not know. He remains ignorant of his own ignorance because he does not question his beliefs and ethical conclusions. Unfortunately, this lesson is learned principally by the reader of the dialogue and not by Euthyphro himself who hurries off at the end of the dialogue to proceed with his prosecution. Socrates will illustrate this lesson in many of his interactions with his interlocutors: the capacity to recognize that one does not know something can be ethically paramount and, further, one can only arrive at this recognition through questioning one's own and others' ethical conclusions.

A second lesson from Socrates regarding responsibility concerns the importance of ethical reasoning being made in personal, reflexive terms rather than in terms of an abstract ethical principle. For instance, when *I* reflect on different ways of acting, it is essential to recognize that *I* am the actor and that the actions are *mine*. Rather than thinking about maximizing the good, or following the categorical imperative, I ought to consider what it means for me to engage in a particular act. This personalization of the context in which one acts is what makes this an ethic of *responsibility*. The individual takes ownership of one's moral agency and understands the ethical significance of one's decisions and actions.

Socrates illustrates this way of thinking by telling the story of his refusal to help kill Leon of Salamis.

> When the oligarchy was established, the Thirty[16] summoned me to the Hall, along with four others, and ordered us to bring Leon from Salamis, that he might be executed. They gave many such orders to many people, in order to implicate as many as possible in their guilt. Then I showed again, not in words but in action, that, if it were not vulgar to say so, death is something I couldn't care less about, but that my whole concern is not to do anything unjust or impious. That government, powerful as it was, did not frighten me into wrongdoing. When we left the Hall, the other four went to Salamis and brought in Leon, but I went home.
>
> (Plato 1997b, 32c-e)

Socrates could have appealed to consequentialism to justify his participation in the murder.[17] He could have appealed to another abstract moral principle to defend his decision not to participate. However, what stands out is his personal concern and his decision simply to go home despite the danger to his life. Readers can locate the crux of individual responsibility here as a perspective that reflects one's understanding that one's decisions and actions are importantly one's own. This fact though simple, is not simplistic, and it is often overlooked.

These lessons concerning a cultivation of individual responsibility through questioning and through personalizing the context within which one acts, are taken up later in the history of philosophy by Søren Kierkegaard. Particularly, in *Fear and Trembling*, one of Kierkegaard's pseudonyms, Johannes de

[16] This refers to the oligarchy that was established after the final defeat of Athens in the Peloponnesian war in 404 B.C.

[17] See Finnis 1983, pp.112-120 for a discussion of this story and consequentialism.

Silentio, writes that "only the one who works gets bread" (Kierkegaard, 1983, III: 79). As we will come to understand, this phrase refers to working towards an individualized and personal reflection upon one's ethical abilities and responsibilities by questioning the soundness and validity of even seemingly self-evident ethical values. Kierkegaard illustrates that a mere acceptance of the given social ethic of one's current cultural milieu is contrary to this end. The individual must challenge even the most apparently evident norms in order to understand them and appreciate them in a personal way. As we will see in the following analysis, Kierkegaard's retelling of the biblical story of Abraham's attempted sacrifice of his son Isaac can be understood as an allegory that describes this process of challenging and questioning ethical norms. The pivotal role this story plays in Kierkegaard's philosophy is not meant as an endorsement of blind obedience to God's will and command, but rather as an urging to question or suspend even those ethical norms as obvious to us as the moratorium on filicide. This "teleological suspension of the ethical" (Kierkegaard, 1983, III: 104) as Kierkegaard calls it, allows for the individual to experience ethical values not as something readily present in and provided by society, but as distinct normative reasons endorsed through personal reflection. This process of ethical reflection and questioning constitutes an ethic of responsibility.[18]

The process by which one questions the accepted values of one's society is difficult. Kierkegaard tells us that one must cultivate a "passionate inwardness" in order to define oneself rather than let oneself be defined by society or culture. For Kierkegaard, when society or culture define one's sense of identity, especially one's ethical identity, then, in his terms, one is not yet a "self". One becomes a self, that is, becomes a responsible individual, through the difficult process of rule- and value-questioning and self-reflection. This passionate inwardness requires one to detach oneself from one's social role and the social ethic that defines it. This detachment requires "sacrificing," or questioning, those values and even those relationships that have the greatest significance for the person. The biblical figure of Abraham as he is portrayed in *Fear and Trembling* engages in this difficult process.[19]

To see this, it is important to note that *Fear and Trembling* consists of multiple retellings and interpretations of the story of Abraham going to Mount Moriah to sacrifice his son Isaac. In so doing, the author Silentio communicates to the reader that the story calls for interpretation and reinterpretation and ought not be taken literally. Taken literally, Abraham is simply someone who follows his faith blindly. Interpreted allegorically, the story of Abraham can tell us about questioning the given values of society, "the social ethic", in order to cultivate individual moral responsibility.

Second, we learn about the nature of Abraham's suspension of the social ethic through Silentio's description of two kinds of knight: the knight of infinite resignation and the knight of faith. Before one can become the knight of faith one must become the knight of infinite resignation (Kierkegaard, 1983, III: 88). This resignation is what Silentio refers to as the "teleological suspension of the ethical" (Kierkegaard, 1983, III: 104). By "ethical" Silentio means the given ethical order of society. He does not call for an abandonment of all morality. This suspension is symbolized by Abraham's attempted sacrifice of Isaac. The resignation is the questioning (or sacrifice) of everything one has been told or taught is good and right, including filicide.

One must question everything, and yet the act of doing so is immensely difficult. Now we are in the position to understand Silentio's assertion referred to earlier that "only the one who works gets bread" (Kierkegaard, 1983, III: 79). After one question the ethical order of society in its entirety, one experiences "fear and trembling" due to the realization that one is personally responsible for one's

[18] For a detailed analysis of this idea throughout Kierkegaard's works, see Goldberg 2017.

[19] Interestingly, Kierkegaard also describes Socrates as a paradigm of one who questions social ethics and engages in self-reflection.

own morality. The individual can no longer rely on the social ethic to define one's responsibility. One must cultivate it oneself.

To accomplish this task on a personal level, Abraham must comprehend what is ethically required of him. By suspending the ethical and defining responsibility through one's personal commitments, one becomes a responsible individual rather than a simple rule-follower. "Johannes de Silentio shakes us from the idea that the key to moral or religious earnestness is in the public sphere where we discover and can then conform to a lucid list of rules" (Mooney, 1996, p. 51). Abraham has responsibilities to his son, not because the role of father has been defined for him by society, but because he questioned this definition and "returned" with Isaac. The father/son relationship has not been destroyed. The message of the allegory is that the relationship has been redefined.

Considering Kierkegaard's retelling of the Abraham and Isaac story in this way, we encounter a compelling way in which to understand ethical responsibility. It cannot be cultivated by the simple following of abstract ethical principles. Rather, it requires questioning values, even those values that seem self-evident, in order to establish a personal relationship with ethical values, as well as one's actions and decision making.

Both of these lessons from the history of philosophy support the claim that questioning ethical values is essential to developing individual responsibility. They both illustrate that adopting and obediently following an ethical rule, even those rules that are ethically justifiable such as those that constitute consequentialism or Kantian deontology, will not elicit in the individual an ability to cultivate a personalized ethic of responsibility—an understanding and acceptance of oneself as a moral agent in addition to ethical ownership of one's decisions and actions. Both Socrates and Kierkegaard's Abraham help illustrate that the moral landscape is filled with perplexing issues making it difficult to navigate. Rather than approaching these difficulties with a ready-made principle purporting to resolve questions before we have identified them, we ought to embrace our and others' ignorance, cultivate our ability to question the roots of this ignorance, and attempt to identify what a particular value means generally, means to society and means to us as moral agents.

It is critical to note that the activity of questioning is not blind. It is an informed questioning brought on by a close familiarity with ethical theories. As Socrates and Kierkegaard illustrate, individual responsibility is cultivated through awareness of one's own and others' ignorance and a commitment to questioning ethical values and principles, even those that seem self-evident. To be able to do so sincerely and earnestly, one must first understand ethical theories, why they are reasonable and credible, but also why each of them faces criticisms and counterexamples. For this reason, the foregoing scholarship in space ethics that focuses upon applying consequentialism, Kantian deontology or virtue ethics to the context of human space activities is indispensable to the process of cultivating individual responsibility in this domain.

QUESTIONING AN ETHIC OF RESPONSIBILITY

As Socrates and Kierkegaard have exemplified, ethical inquiry is a continuous activity. It is appropriate, therefore, to probe further into our discussion on questioning and its connection to the ethic of responsibility.

Aren't there Already Existing Contemporary Versions of this Approach?

The most developed philosophical treatments concerning an ethic of responsibility appeared originally in feminist philosophical literature (e.g., Held 1987, Meyers 1989, Ruddick 1989, Whitbeck 1984), but

have since been taken up by non-feminist philosophers as well including Christopher Gowans (1994) and Larry May (2006). These influential theories have helped to develop philosophical thinking concerning the source and content of moral responsibility. Although the current approach is clearly influenced by these earlier theories, it is also significantly different in two ways. First, the current approach is not limited to human-human interaction, and second, it is includes epistemic responsibility as a kind of ethical responsibility.

Although the above-mentioned authors and their theories possess significant differences, one insight that they share is the importance of an individual's relationships with others. What is similar between these approaches and the current one is the conclusion that one cannot simply follow an abstract ethical rule to be an ethically responsible person. According to Gowans, we can see this fact when we reflect upon what is involved in our particular, concrete relationships with persons with whom we are to a greater or lesser extent, and in various ways, intimate, especially relations of kinship, friendship, and love. It is mainly in the context of such relationships that we first come to employ and understand moral considerations (Gowans, 1994, 122). Although all our responsibilities will not be found in intimate relationships, these relationships serve as paradigms on the basis of which responsibilities in other contexts may also be understood. We all suppose that we have specific responsibilities to those persons with whom we are so related.

Larry May likewise rejects traditional ethical theories as sufficient guides for responsible action, and states that one ought to focus on the following elements of human interaction:

> 1. A responsiveness to those whom we could help, especially concerning those who are in relationships with us or toward whom we have taken on a certain role;
> 2. a sensitivity to the peculiarities of a person's concrete circumstances and contexts;
> 3. a motivation to respond to another which grows out of the needs of others, especially those who depend on us;
> 4. a wide discretion concerning what is required to be a responsible person, rather than an emphasis on keeping an abstract commandment or rule;
> 5. a respect for the legitimacy of emotions as a source of moral knowledge, and especially for the feelings of guilt, shame, and remorse that are central to people's actual moral experiences;
> 6. a sense of who we are as responsible people that is tied more to who we are, and what we can do, than to what we have done.
>
> (May 2006, 402)

May argues that these guidelines are more reflective of ethical decision making and action than consequentialism or Kantian deontology precisely because the guidelines take into consideration our actual moral experiences and personal relationships. According to May, responsibility is constituted by reflection and sensitivity to these six points.

Nevertheless, Gowans' and May's accounts are too limited for the responsibility needs of the space context. Rather, the ethic of responsibility described in this essay is meant to be applicable to human/animal/environment/alien/anything interaction. The cultivation of responsibility through questioning values can and ought to occur regardless of the situation, but especially in situations with significant unknowns such as that of human space activities. Secondly, the current ethic of responsibility is not only about interaction with others. It also concerns self-reflection. It is a reflective,

and oftentimes reflexive, act that has as its objective the epistemic goal of recognizing one's own and others' ignorance. As Socrates clearly illustrates, this epistemic achievement is itself an ethical act.

Isn't this Ethic of Responsibility Derivative of Other Ethical Theories, especially Virtue Theory?

The conceptual and practical reliance on justified ethical theories is a positive aspect and is essential to a cultivation of responsibility as it is conceived of in this essay. As both Socrates and Kierkegaard's Abraham have revealed, one must first have a serious understanding of ethical theories such as consequentialism and Kantian deontology in order to question and challenge them appropriately. These ethical theories are insightful and possess many correct conclusions about right and good actions. However, they are also limited in certain ways and are subject to ethically significant criticisms. In order to continue to pursue these criticisms and to identify their replies and newly developed criticisms, one must have a fine-grained understanding of the theories. In this sense, the ethic of responsibility is derivative of traditional ethical theories.

Indeed, the practical wisdom that constitutes virtue is often defined as the wisdom to establish the correct moral principle for a given case. As revealed above, the purpose of the cultivation of virtue is moral expertise. In fact, for traditional virtue theorists, virtue is synonymous with moral expertise, and expertise is constituted by knowledge and understanding (Vallor 2016, 26). Again, the ethic of responsibility is heavily influenced by virtue theory, but it contains a critical difference. At the core of the ethic of responsibility is the recognition that we cannot be certain what the correct moral principle for a given case is, and yet we must decide and act anyway. [20] The ethic of responsibility presupposes knowledge and understanding of the limits of one's knowledge and understanding. It is precisely the awareness of these limits and the need to continue questioning that makes the ethic of responsibility essential for circumstances of significant unknowns such as the context of human activities in space.

Does Adopting an Ethic of Responsibility Approach Guarantee Ethically Good Decision Making and Action?

Unfortunately, not. One can question and challenge ethical theories and societal norms and draw the conclusion that money or power, or some other non-ethical desire, are more important than acting ethically and responsibly. In fact, after becoming aware of the significant challenges faced by each traditional ethical theory, one might conclude that there are no objective truths about ethics and that ethical values are merely societal obstacles that the clever can navigate in order to achieve their preferred individual ends. This risk is one of the reasons that Kierkegaard's Johannes de Silentio writes of the sincere struggle accompanying the challenging of ethical norms. Once one questions the ethical validity of the prohibition against filicide, one might conclude that the prohibition is invalid or unjustified. In the fable, there is no guarantee that Abraham returns from Mt. Moriah with Isaac alive. However, because he does return with Isaac alive, he sees his ethical responsibility in a new and personal manner. It is owing to this very risk, the possibility of returning from the plane of questioning as an ethical villain, that makes the cultivation of responsibility genuinely difficult, but also ethically rewarding.

[20] For example, this methodology and approach have been successfully carried out by the author in the D4FLY project, which is a project under European Union's Horizon 2020 research and innovation programme, grant agreement No 833704. The project designs new technologies for border security, and, given the novelty of the technologies, also involves several ethical unknowns.

Why is this Approach Appropriate for the Context of Human Activities in Space?

I have indicated throughout this essay that the ethic of responsibility approach ought to be utilized in contexts with significant unknowns. The ethics of responsibility is particularly applicable to human activities in space, where we are dealing with significant unknowns to an even greater extent than on earth.

For example, although the attempt at estimating the value of the consequences of human activities in space will continue to be ethically worthwhile, the unknowable elements of the consequences of our actions in space make a consequentialist approach helpful but woefully incomplete. To any consequentialist approach will be needed earnest critique of the consequentialist method conclusions. For example, one might view through a consequentialist lens the questions whether it would be ethically justified to mine Mars for minerals to be used as resources on Earth or whether it would be ethically justified to look for signs of previous life such as fossilized microbes. On the one hand, the activity of mining will disturb the Martian environment never before touched by any human. Perhaps there is some value in such "pristine" environments. On the other hand, humans are in need of new and renewable resources to support their lives and activities on Earth. Furthermore, the discovery of extra-terrestrial life, even in microbial form, would be an immensely significant scientific discovery and could provide critical understanding about Mars and, since Earth and Mars had relevantly similar environments until about 10 billion years ago, provide invaluable information about how life on Earth emerged and evolved.

The ethic of responsibility does not prompt questions about how to adjudicate these conflicting consequences and their values or to help us imagine unforeseen consequences. These considerations are essential, but still belong within the standard consequentialist framework. The ethic of responsibility goes beyond this level of reflection and elicits questions about the value of thinking about consequences at all. It also shifts the focus from the non-personal value belonging to states of affairs to contemplating the personal manner in which value or disvalue is created through an individual person's, including my own, decisions and actions. It prompts us to ask about the impossibility of making an ethically good decision given our ignorance under particular circumstances and yet, we must nevertheless make some decision and take some action. One cannot simply take refuge in an abstract ethical principle. How does one, how do I, take responsibility for my decisions and actions? How do I make amends if I inadvertently violate the ethical values I want to protect and promote? What if I cannot repair the violations? The purpose of raising these questions is not to resolve them, as would be the purpose under other ethical frameworks. Rather, the activity of raising questions while acknowledging that they cannot be satisfactorily answered is the core of an ethic of responsibility. Given the vast unknowns surrounding and arising from human activities in space, the continuous activity of raising ethical questions is paramount.

CONCLUSION

The context of human activities in space raises significant ethical questions. Given the pervasive unknowns in this context, it is impossible to answer ethical questions without significant doubt as to whether the answers are sufficient and accurate. However, this doubt should not be seen in a negative light. Recognizing that it exists and understanding that it exists owing to our own and others' ignorance about how to protect and promote ethical values in this context, or even which ethical values are applicable to the context of space, is an essential step in the cultivation of an ethic of responsibility. As illustrated by Socrates and Kierkegaard, ethical responsibility requires questioning and challenging the ethical status quo in order to identify gaps in understanding as well as to personalize ethically relevant action. Although this ethic of responsibility ought to be cultivated in any context, especially in contexts

with significant unknowns, it is particularly relevant for the context of human activities in space because in space even the quality and quantity of the unknowns are still unknown. Rather than engaging in these activities with supposed ethical expertise, it is prudent to engage in them with awareness of our ethical ignorance.

KEYWORDS: Space ethics, Responsibility, Ethical theory, Ethical Inquiry.

REFERENCES

Arnauld, J. (2011). *Icarus' Second Chance: The Basis and Perspectives of Space Ethics*. Vienna: Springer.

Baum, Seth. (2016). The Ethics of Outer Space: A Consequentialist Perspective. In: J. Schwartz and T. Milligan (Eds.), *The Ethics of Space Exploration* (pp. 109-123). Berlin: Springer.

Daly, E. & Frodeman, R. (2008). Separated at Birth, Signs of Rapprochement: Environmental Ethics and Space Exploration. *Ethics and the Environment*, 13(1), 135-151.

Drengson, A. 1981. The Virtue of Socratic Ignorance. *American Philosophical Quarterly*, 18(3), 237-242.

Finnis, John. (1983). *Moral Absolutes*. Oxford: Clarendon Press.

Goldberg, Z. (2017). Moral Innocence as the Negative Counterpart to Moral Maturity. In: Dodd, E. & Findley, C. (Eds.), *Innocence Uncovered: Literary and Theological Perspectives* (pp.167-182) London: Routledge.

Green, B.P. (2014). Ethical Approaches to Astrobiology and Space Exploration: Comparing Kant, Mill, and Aristotle. *Special Issue "Space Exploration and ET: Who Goes There?" Ethics: Contemporary Issues* (2)1, 29-44.

Gohd, Chelsea. (2021). Elon Musk reminds us that 'a bunch of people will probably die' going to Mars. Retrieved from https://www.space.com/elon-musk-mars-spacex-risks-astronauts-die May 3, 2021.

Gowans, C. (1994). *Innocence Lost: An Examination of Inescapable Wrongdoing*. Oxford: Oxford University Press.

Held, V. (1987). Non-contractual society: A feminist view. *Canadian Journal of Philosophy* Supplementary 13: 111–137

MacArthur, D., Boran, I. (2004). Agent-Centered Restrictions and the Ethics of Space Exploration. *Journal of Social Philosophy* 35(1), 148-163.

Matignon, L. (2019). Space Tourism Legal Aspects. Retrieved from https://www.spacelegalissues.com/space-law-space-tourism-legal-aspects/

May, L. (2006). Social Responsibility. *Midwest Studies in Philosophy* 20, 400-415.

McElreath, S. (2017). Contemporary Virtue Ethics and Action Guiding Objections. *South African Journal of Philosophy* (37)1, 69-79.

Meyers, D. (1989). *Self & Society und Personal Choice*. New York: Columbia Press.

Mooney, E. (1996). *Selves in Discord and Resolve: Kierkegaard's Moral-Religious Psychology from* "Either/Or" *to* "Sickness Unto Death". New York: Routledge.

Plato (1997a). Euthyphro. In John M. Coooper (ed.), *Plato: Complete Works* (pp. 1-16). Cambridge: Hackett Publishing Company.

Plato (1997b). Apology. In John M. Coooper (ed.), *Plato: Complete Works* (pp. 17-36). Cambridge: Hackett Publishing Company.

Ruddick, S. (1989). *Maternal Thinking: Toward a Politics of Peace*. Beacon Press, Boston, MA.

Schwartz, J., Milligan, T. (2016). *The Ethics of Space Exploration*. Dordrecht: Springer.

Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford: Oxford University Press.

Whitbeck, C. (1984). A Different Reality: Feminist Ontology. In: Gould, C. (ed.) *Beyond Domination*. Totowa, N.J: Rowman & Allanheld.

TOWARDS IMPROVING THE DECISION-MAKING PROCESS OF ARTIFICIAL INTELLIGENCE DEVICES IN SITUATIONS OF MORAL DILEMMAS

Damian Węgrzyn

Polish-Japanese Academy of Information Technology (Poland)

damian@wegrzyn.info

ABSTRACT

Systems using Artificial Intelligence (AI) are created by humans to achieve specific goals. As autonomous decision-making increases, one of the most important considerations is the need to rethink its responsibility. AI's decisions can affect many key aspects of human life. The main issue that arises in the discussion about the usage of AI is the ethical nature of decisions made by AI. Moreover, in some situations these decisions are related to moral dilemmas. This paper deals with problems of moral dilemmas and analyses the moral status of AI devices. As a suggestion to improve the decision-making process in situations of moral dilemmas, the author proposes to apply the fuzzy logic theory. This solution is already used in making choices by AI systems but so far it is not applied to ethical choices in crucial situations of moral dilemmas.

INTRODUCTION

More than 60 years ago, the term Artificial Intelligence (AI) was defined as the tasks performed by a device previously claimed to require human intelligence (McCarthy et al., 1959). Nowadays, the authors attribute a wider area of designation to this concept. The definitions touch upon new skills, such as the ability to flexibly adapt, to learn or even to make decisions based on newly acquired data (Kaplan & Haenlein, 2019). In addition, AI is given the task of adapting through a learning process leading to the ability to sense, reason and act in the most efficient way possible (Tørresen, 2020). After all, it is described as a vague concept with many open questions remaining.

AI is already involved in our everyday life. It has an impact on such important issues as safety, human life and health (Stone et al., 2016). AI can be found in areas such as biomedicine, education, finance, energy, law, space exploration, etc. Good examples of modern applications of AI are aircraft autopilots, where in critical situations the pilot can take control of the machine using his or her experience and acquired skills. In many cases, devices with AI even replace people in making various decisions, such as driving cars, making credit decisions or interpreting medical research results. The rapid development of devices with AI and their entry into everyday life requires them to make decisions. AI systems that make decisions in various areas can affect many key aspects of human life. This raises many questions related to the ethics of decision-making by AI devices. Researchers want to ensure that these systems are ethical, but this is not easy to achieve. Still, the system developers should enable the AI systems to make ethical decisions (Dennis et al., 2015).

Understanding the reasons behind the choices made by modern AI machines is either difficult or sometimes even impossible. This is due to the complexity of the processes that constitute the final choice, such as deep learning using Artificial Neural Networks (ANN). Therefore, there is a need to urgently look at particularly crucial decisions. Teaching the machines morality is undoubtedly a difficult task. Some scenarios cannot be predicted or programmed. Furthermore, there are situations - the so-

called moral dilemmas - in which even a man has doubts about what to choose. AI software architecture uses measurable metrics that are not designed for objective moral evaluation. This is because morality is a concept that includes aspects that are not measurable. While the distinction between what is good and evil has at its base an arbitrary or customary set of norms, the definition of many acts is already burdened with subjectivism. In situations of moral dilemma, making choices is often determined by feelings, benefits, or internal prejudices. This is undoubtedly not the case in AI systems. It is known that a machine can only be taught to understand concepts properly if the engineers who design it have a precise definition of the concept. In many cases, the solution is the optimization of decisions, although in real situations it does not always work, because it leads towards the principle of equality and not necessarily justice. In some situations, the usage of rigid algorithms has resulted in discrimination, prejudice or inappropriate choices (Bartneck et al., 2018). As of today, we are unable to teach AI to make fair choices because we do not have an unchanging evaluation within this basic concept of ethics that is out of context or person. In extreme cases, the use of the choices optimization process, without constant analysis and relation to the reality of those choices, may lead to making biased or wrong choices.

Making ethical decisions is a controversial issue, too. When we consider extreme moral dilemmas, in which even people have doubts about the final decisions, we are faced with a problem that is impossible to solve algorithmically, i.e., choosing the lesser evil or the greater good. Values assumed to be immeasurable are elusive for modern technological solutions.

PROBLEMS OF MORAL DILEMMAS

The fundamental problem with ethical dilemmas is whether they exist. Opinions are divided on this point (Holbo, 2002). This paper tends to argue for the existence of moral dilemmas and assumes that the solutions available to the subject are of equal value: none prevails over the other. Moreover, the author supports the thesis that there are no ideal moral theories that would allow one to make ethical choices in every situation. This paper treats a moral dilemma in the strict philosophical meaning. So determined moral dilemma fulfils all the following conditions:

1) an agent is faced with a choice situation between at least two solutions to the problem,

2) each of the solutions to the problem can be chosen by the agent,

3) all solutions are not identical and contradict each other,

4) none of the solutions is subordinate or superior to the other,

5) the agent should select; if there is no choice, one of the available solutions,

6) the agent may choose *n-1* solutions among *n* available solutions,

7) any choice among the available solutions or no choice made brings with it immoral consequences.

Types of moral dilemmas are related to the assumptions made. In general, a distinction can be made between solvable and unsolvable dilemmas. A deeper division involves epistemic dilemmas (one of the solutions takes precedence in a situation) and ontological dilemmas (there are no superior solutions) (McConnell, 2018). This article deals with unsolvable, ontological and symmetrical dilemmas, in the case of which the same moral precept gives rise to conflicting obligations (Sinnott-Armstrong, 1988).

The analysis of the behaviour of AI devices in unforeseen situations introduces uncertainty and imprecision in decision-making. When it comes to situations of moral dilemmas, some of them can be predicted and general scenarios of behaviour or decisions can be prepared for them. Unfortunately,

there is a wide spectrum of unforeseeable events in which AI systems will have to make a choice. Then, when deciding or judging, they must be based on the ingrained basic principles of ethics and the data collected so far. It is not known if this ultimately suffices to make choices that can be considered moral.

Defining moral values is a challenge that mankind has grappled with throughout its history. Policymakers and engineers should have methods which allow implementing ethical standards that bring them closer to quantifying ethical values. Finally, let us remember that AI devices are made by people who are subjective and biased in their judgments. By creating ways of ethical choices in situations of dilemma, it is possible to reproduce human faults in AI systems, because ultimately it is a human being who creates AI systems' behaviour and decisions. In this context, the value and cost of both subjective and objective decision-making must be considered. It could be argued that both are needed, albeit to a different extent, to ensure control and balance. Since AI cannot adequately deal with subjectivism, it cannot be deemed to be ethical. However, the actions that will be taken by the AI system are subject to an ethical evaluation. Furthermore, subjectivism is a problem that results directly from human nature and is an inherent factor of choices. The right way to reduce subjectivism and at the same time increase objectivity is to expand the set of choices made in similar situations by people. Crowdsourcing is currently used for this purpose, in which it is assumed that individuals have good intentions and make moral choices. While in known situations it can be inferred with the use of this method in a highly objective manner, the problem remains in new, dilemmatic choices.

Another known problem in evaluating choices is the situational, activity, personal and intentional context. Depending on the context, people evaluate specific facts and make decisions. In the case of an intentional context, morality imposes the choice of good or less evil, in accordance with the current state of knowledge of the decision-maker. The situational and activity context indirectly affects the decisions made, as they may have the so-called mitigating effects. In the case of the personal context, the problem becomes multidimensional. On the one hand, current standards and norms do not allow AI systems to make significant choices based on personality attributes (Di Fabio et al., 2017). On the other hand, there may be situations where the lesser evil is chosen based on the personal context, such as in the trolley problem (Thomson, 1985). Moreover, the context often depends on other phenomena or may even constitute a tangle of events. In such a situation, it is almost impossible to predict the situation or program the scenario.

An important problem with morality in general is the imprecision in defining and assessing moral attitudes. This is because in the final assessment, the problems mentioned so far, such as subjectivism, context or a combination of events, participate to a different extent. In such complex situations people find it difficult to make an honest judgment. The variety of assessments may result, inter alia, from the fact that each of the factors constituting a judgment or choice receives a different weight.

MORAL STATUS OF ARTIFICIAL INTELLIGENCE DEVICES

Nowadays, developers provide AI devices with specific decision rules in situations of moral dilemmas. This requires establishing and defining ethical norms of behaviour in specific difficult situations. The mere implementation of current available and determined indicators allowing to define ethical values is not sufficient for AI systems to decide ethically in all situations. To train ANN models, a large set of unambiguous examples should be collected. For the model to be properly trained and the output to be predictable, as many human judgments as possible must be gathered. By design, this involves showing AI devices clear answers and decision rules to the potential ethical dilemmas they may encounter. It comes down to establishing the most ethical course of action in a difficult situation. Only in this way is it possible to increase the objectivity of decisions due to the variety of situations.

Crowdsourcing is used for such purposes, especially when designing autonomous AI devices, assuming no one is deliberately suggesting the wrong solution. However, in unpredictable new cases AI systems are on their own and have to make choices.

There are exceptions to the rule, which represent the deliberate unethical (in the usual sense) usage of AI devices. One of them is the practice of using IT systems by modern developers for such purposes as lethal autonomous weapons systems (LAWS), drones - killers. The development of information technology proves that war is an important engine of technological progress. It is for military needs that new projects and technologies are constantly being developed, which are ethically controversial.

In terms of the moral status of AI, it is widely assumed that modern AI systems do not have moral status - they are amoral (Bostrom & Yudkowsky, 2014). To categorize a being as having a moral status requires it to belong to a kind that has a sense of sensitivity or reason in the normal way. This can only be done concerning an entity for whom there is no doubt of having an independent moral status (Warren, 1997). If AI devices as self-learning and self-modifying beings will have a sense like the human mind, special attention should be paid to its initial state, as this may have permanent effects and negatively affect its further, ethical self-development, and thus cognitive functions of good and evil (Omohundro, 2008). It is known from the history of ethics that in different periods of mankind the concept of ethics and basic ethical norms have evolved. Once upon a time, slavery was the norm. Quite recently, in the 19th century, women were generally denied the right to vote, as were people of another colour of skin. These days this is categorized as discrimination or even racism (Bostrom & Yudkowsky, 2014). There is a chance that with such a dynamic technological development, AI systems will acquire a moral status, and even become an interpretation of ethics - as an entity with the ability to objectively judge, better than one person, e.g. a judge issuing a judgment in a case.

There are some complications in the direction of AI algorithms towards human thinking. They can fulfil certain social roles, which implies new design requirements such as transparency and predictability. Sufficiently broadly targeted AI machines can operate in unpredictable contexts, requiring security and engineering to incorporate ethical aspects.

Fundamental current issues connected with AI device ethics are transparency, privacy, and awareness of AI (Green, 2017). The decision-making process of AI systems with the complex structure of ANNs is not transparent. Therefore, it is not known on what basis the machine made a specific decision and is not able to explain it. This is known as the black box problem (Winfield, 2017). Nonetheless, AI system designers should make AI device decisions more transparent in an ethical context. Full transparency cannot be ensured, but there is room for greater transparency on how to get closer to quantifying ethical values in programming and determine the choices ultimately made by AI.

It is unacceptable to justify AI's incorrect behaviour by doing nothing about it. By detailing the decision possibilities that AI can make, it allows us to avoid uncontrolled and dangerous decisions of AI systems, especially in situations of ethical dilemmas. Creating algorithms that define a set of ethical values, which are the premises for making AI decisions, will also serve to avoid significant harm. Since people learn moral principles, it must be assumed that systems with AI can also follow unethical paths unconsciously and unintentionally. This requires engineers to constantly improve their moral definition and try to quantify it, which is extremely difficult. In the history of ethics, researchers have attempted to quantify non-measurable attributes to determine the moral status of a given act. A notable example is Bentham's ethical account (Brunius, 1958). This theoretical algorithm of human action describes it as the pursuit of pleasure and avoidance of pain - in line with the hedonistic postulates. The calculations were based on a vector of seven variables (intensity, duration, certainty, speed of occurrence, efficiency, purity and scope). Bentham also defined many kinds of pleasure and distress that a person chooses in certain situations. To evaluate the moral act, the function of the

amount of pleasure or pain induced was used. The proposed measurement method significantly simplified the concept of human nature: it reduced the multidimensional complexity of human action to a binary system in known situations. Such a simplified categorization of two values included the hierarchy of not only ethical values but also aesthetic, cognitive and material ones.

AI devices are incapable of moral behaviour. It is their creators who must lay the foundations for their understanding of morality - how to analyse it discreetly. The history of ethics shows that it is not easy to define and quantify such concepts. Ultimately, it is impossible to implement the entire morality in the behaviour of AI devices in a situation where there are no unambiguous and measurable attributes of this issue. Nevertheless, apart from the implementation of ethical behaviour, we leave AI the right to decide also in critical situations.

DISCUSSION ABOUT STANDARDS, NORMS AND RESPONSIBILITIES

To integrate moral or social values with the technological development of AI at all its stages: design, analysis, construction, implementation and evaluation, well-thought-out standards, methods and algorithms are necessary. The idea behind these recommendations is that such devices should be able to make ethical decisions based on a general ethical framework. There is a growing desire among AI engineers for these technologies to be fair and ethical. Standards and norms are usually developed by experts in many areas, which guarantees that it will be an ethically acceptable process.

Recommendations for AI developers

The area of computer ethics' interest is the ethical guidelines for machines with AI. Since ethics should constitute the basis of standards, it is impossible to omit them in this discussion. The list of the current official recommendations in the literature that deal with AI's ethical dilemmas includes the standards of the European Union's Roboethics Special Interest Group (Veruggio, 2006), South Korean Robot Ethics Charter (Korea's Ministry, 2012), reports of German Ethics Commission (Di Fabio et al., 2017), the BS8611 standard by British Standards Institute (British Standards Institute, 2016), and IEEE's Ethically Aligned Design (IEEE, 2019).

The first three basic principles of ethics in the process of creating machines were introduced by Isaac Asimov in the 1940s, known as Asimov's Laws (Asimov, 1942). The first law is principal and covers the protection of human health and life through devices. The other two laws are only a supplement to the first, as they regulate the behaviour of robots concerning the implementation of human commands and the validity of the device's survival. Additionally, the third law tracks the decisions AI devices make - a machine cannot put its existence ahead of human health.

The European Union has secured the ethics of AI machine development with the establishment of the Roboethics Special Interest Group. Article number 1.1 of E.U. Standards (Veruggio, 2006) is concerned with the safety and autonomy of robots. It recommends that the robot have its operators who should be able to limit the autonomy of devices in situations, where their behaviour cannot be guaranteed. This also includes decisions made in situations of ethical dilemmas. This standard should be implemented in all types of robots.

A good example of a standard that describes recommendations for decisions made by AI devices is the South Korean Robot Ethics Charter (Korea's Ministry, 2012), which was established in 2006 and updated in 2012. In the part describing manufacturing standards, it is recommended that in critical situations, AI devices should be prepared for human control. Robot manufacturers should be mindful

of minimizing the risk of death or injury to the user, as well as keeping the community safe. In addition, the topic of antisocial and sociopathic behaviour by robots is discussed to minimize the risk of psychological injury to humans. In the second part, dealing with the rights and obligations of users, the standard guarantees them the right to use the robot without risk or fear of physical or mental harm and the right to take control of the robot. The Charter also gives users the right to use the robot in any way if it is fair and within the law. Separately, it is mentioned that the user is not allowed to use the robot in a way that causes physical or mental harm. In the third part, concerning the rights and obligations of the robot, a clause has been added that the AI device cannot deceive a human, and therefore its decision-making should be clear, obvious and transparent in this aspect.

The German Ethics Commission took up the issue of moral dilemmas in the decisions of AI devices. The guidelines were published in 2017 in the report for Automated and Connected Driving (Di Fabio et al., 2017). Clause 5 recommends that AI devices should be designed to avoid critical situations. The authors consider moral dilemmas as situations, in which a machine must choose between two unethical outputs, between which there is no compromise. Thus, it is proposed to continuously develop the entire spectrum of technological options that will allow for anticipation and decision-making with the least possible risk to humans, thereby increasing safety. If, on the other hand, there is a critical situation that cannot be avoided by using available technological solutions, first of all human life should be protected. Therefore, the AI systems must be programmed to accept damage caused to animals or property in a dilemma situation, leading to the risk of human health and life. Extreme dilemma decisions, in which there is a choice between one human life and another, depend on the specific situation and cannot be uniquely standardized or programmed. There is no standardization of the effect's assessment of decisions, which would be equivalent to a person's moral capacity to make judgments under certain circumstances and historical data. The publication emphasizes that transforming such processes into abstract or general ex-ante evaluations in the form of appropriate programming activities is extremely difficult. For this reason, it is recommended that independent institutions systematically process the lessons learned from the behaviour of AI devices. Decision-making in moral dilemmas cannot be based on personal characteristics such as gender or age. AI programming should be based on the principle of reducing the number of injuries. AI systems also cannot make decisions related to the sacrifice of the other party. Clause 16 differentiates between a fully autonomous system and a system that can be nullified. The second type should be designed in a way that allows for an unambiguous assignment of responsibility: whether it is on the side of the AI system or the side of the user. Clause 18 allows self-learning systems to be implemented only when the security requirements are met and without questioning fundamental ethical principles. Connectivity to the scenario databases is also acceptable if there is a security benefit. However, it is recommended to develop an appropriate standard, including acceptance tests, based on a catalogue of scenarios. Ultimately, in critical situations, the AI system must be able to enter the so-called safe condition, without external human assistance.

British Standards Institute published in 2016 the BS 8611 standard (British Standards Institute, 2016) containing guidelines for the safe design and use of robots. This standard guides to help eliminate or reduce the risk of ethical risks associated with the use of robots. The analysis was based on the standards related to the risk assessment of machines, as well as risk reduction and management. The standard defines various terms, especially ethical harm, ethical threat and ethical risk, which allow for a general understanding of the key and basic principles that determine human behaviour affecting programmed AI devices. A similar approach is presented in the IEEE document published in 2019 (IEEE, 2019). It introduces the vocabulary and models of risk assessment to explain ethical dilemmas.

The responsible innovation

The literature on the ethics of AI devices emphasizes the analysis of responsibility (Dignum, 2017). The authors' conclusions boil down to the issue called responsible innovation (Wong, 2016). This idea assumes that the responsibility for AI machines also rests with manufacturers. This is consistent with the thesis that AI's responsibility is fundamental. P.H. Wong (Wong, 2016) notes that creating AI should be more like raising a child than programming an application.

Responsibility for the development of AI devices is to ensure compliance with basic human principles and values to ensure order and prosperity in a sustainable world. In this context, the creation of AI machines, as an element of responsible innovation, consists of ethics by, in, and for design. V. Dignum (Dignum, 2018) defines ethics by design as integrating the ability to ethically reason in an algorithmic way into the behaviour of AI systems. Ethics in design includes regulatory and technical methods to support the evaluation of the ethical consequences of AI devices that participate in human social structures. Ethics for design assumes a close relationship between developers and users at all life cycles of AI systems in the form of codes of conduct, standards, or certification processes. P. Vamplew et al. (Vamplew et al., 2018) raise issues of legal, ethical and security frameworks that are not sufficient for multi-purpose decision-making by AI systems. The authors propose the paradigm of the multiobjective maximum expected utility. It is based on a combination of vector tools and a non-linear selection of activities, allowing to determine the current effectiveness of the maximum expected utility. T. Arnold and M. Scheutz (Arnold & Scheutz, 2018) propose a scenario generation mechanism that allows to verify the decisions of AI systems in the virtual world to avoid them in the real world. V. Bonnemains et al. (Bonnemains et al., 2018) analyse the ethical reasoning of AI systems. The authors propose an automatic process of judgment of decisions from an ethical point of view, based on models of ethical principles and formal tools describing the situation. To answer a specific ethical dilemma and its moral assessment, the authors use modelling in three ethical areas: utilitarian ethics, deontological ethics and the doctrine of double effect.

Currently, there are many projects in the field of AI ethics development, having a significant impact on the analysis of moral dilemmas of AI devices. They support cooperation in the field of AI ethics (Partnership on AI project) and deal with ethics in autonomous systems (IEEE Ethics in Action in Autonomous and Intelligent Systems project, IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems project). Some projects directly contribute to improving the quality of ethical decisions through crowdsourcing (Moral Machine project).

PROPOSAL OF THE FUZZY LOGIC APPLICATION

On the one hand, ethical standards and recommendations suggest that AI systems should not be guided by individual characteristics, such as age, gender, or physical or mental constitution in critical choices between the preservation of two persons' lives (Di Fabio et al., 2017). On the other hand, there should be a choice of the lesser evil or the greater good. One of the potential possibilities is a hierarchical order of solutions or values. Then the choice seems obvious but it is not reliable. In the case of moral dilemmas, although the available solutions are contradictory, they are nevertheless on the same moral level. The hierarchy of solutions leads to a situation where one of the choices will be morally inferior, which excludes the attributes of a moral dilemma.

In decision-making processes, the ability to deal with uncertainty and imprecision is an important issue and affects the quality of decisions made. Imprecise concepts are attributes of human judgment that are reflected in the process of AI systems programming. Therefore, a mathematical formulation with precise values cannot describe and predict a realistic decision-making process (Xue et al., 2017). To

describe fuzzy attributes, a fuzzy inference system can be used, mapping the relationships between many decision components. In such a process of inference, the use of the fuzzy logic theory may be helpful.

The fuzzy logic between the two extreme states 0 and 1 assumes many intermediate values that determine the extent to which the element belongs to the fuzzy set (Zadeh, 1965). In the discretization of fuzzy concepts, all states should be assigned discrete values. Fuzzy logic is currently used in control systems, evolutionary algorithms and neural networks, based on which it is possible to build decision-making systems that analyse ambiguous or sometimes even contradictory features. Nowadays, the fuzzy logic theory is at the base of the decision-making process. There are examples of AI systems based on fuzzy logic in the literature. For instance, the proposal of a pedestrian recognition model that incorporates fuzzy logic into a multi-agent system, to deal with cognitive behaviours that introduce uncertainty and imprecision in decision-making, confirms the high effectiveness of this method (Anderson and Anderson, 2018; Xue et al., 2017). However, these are not ethical decisions. First of all, these methods are based on various personality models that represent features of human nature. In the case of the decision-making process of AI devices, e.g. autonomous vehicles, this is not allowed due to the requirements of applicable standards. In the case of ethical decisions, sets defining unmeasurable values (e.g., from crowdsourcing) can be used to help describe imprecise ethical concepts, such as evil, good, justice, or freedom but it cannot be the final criterion, due to the subjectivism of individual human assessments, mistakes in the answers given or even the immoral goals of individuals. Still, the main advantage of such a solution is the possibility of modelling ethical behaviour simulating human nature, which is ambiguous and imprecise.

The fuzzy set theory in the decision-making process of AI systems in situations of ethical dilemmas can be used in conjunction with many available decision support methods (Ogryczak, 1997). One possibility is the concept of decision preferences. The concept of the preference relationship is currently the basis for researching decisions of individuals. Direct relationship measurement is a difficult task. Preferences are characterized as a binary relation referring to vectors describing multidimensional objects. In formal terms, preferences are usually a preorder or a linear order, i.e. a reflexive, transitive, and consistent binary relation. The relation of preferences enables the decision-maker to be assigned an individual scale of preferences, on which profiles can be evaluated and choices optimized. The function of assigning a value to individual preferences is an ordering function that introduces an order (Bąk, 2013). At this point, to apply the concept of preferences in the decision-making process in moral dilemmas, the function of belonging to fuzzy sets can be used, which will allow for a more realistic moral evaluation of the choice. Such fuzzy inference also considers the features of the decision-maker and can generate different decision preferences, because fuzzy relationships between decision preferences are determined by the fuzzy inference system. Inference in situations of moral dilemmas is a multi-criteria inference, where different solutions may have different vectors of moral evaluations. In addition, there is no general or a priori formulated function. Therefore, the incomparability of individual solutions in the sense of the model does not mean that they are incomparable or indistinguishable. Sometimes it is assumed that in such a situation the set of solutions to the problem is the whole set of effective solutions (Ogryczak, 1997).

The suggestion of using fuzzy logic supports both the objective and subjective approach, because on the one hand it is based on previously known standards, norms or crowdsourcing, and on the other hand, it analyses the decision-making preferences of the decision-maker concerning the dynamic situation, effects and context. Thus, currently used mathematical decision-making mechanisms can be used in situations of unsolvable and ontological dilemmas.

CONCLUSIONS

Nowadays, IT systems make decisions in various areas of everyday life, or they will do so soon. The presented analysis of the complexity of the choices made by AI devices shows the need to increase human safety and brings AI judgments closer to objectivity and ethical behaviour. Teaching the AI devices morality is undoubtedly a difficult task because of its immeasurable character. This paper sets out a possible direction that integrates fuzzy logic theory into the decision-making process of AI systems where there is uncertainty and imprecision.

This paper contributes to the ongoing debate on the automation of ethical decision-making through AI. It shows the importance of this issue and outlines the direction of further research in the moral dilemmas area. The author indicates the importance and complexity of the problem. The presented deliberation of issues related to moral dilemmas in the area of AI is an incentive for further analysis, research and implementation.

ACKNOWLEDGMENTS

The author would like to show his sincere appreciation for helpful comments and kind suggestions for the manuscript and a great deal of encouragement from Professor Alicja Wieczorkowska of the Polish-Japanese Academy of Information Technology, Poland.

This paper was partially supported by research funds sponsored by the Ministry of Science and Higher Education in Poland.

KEYWORDS: moral dilemmas, AI devices, decision-making process, fuzzy logic.

REFERENCES

Anderson, M. & Anderson, S.L. (2018). GenEth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics*, 9(1), 337-357. https://doi.org/10.1515/pjbr-2018-0024

Arnold, T. & Scheutz, M. (2018). The “big red button” is too late: an alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology*, 20, 59-69. https://doi.org/10.1007/s10676-018-9447-7

Asimov, I. (1942). Runaround. *Astounding Science Fiction*, 29(1). Retrieved from http://www.isfdb.org/cgi-bin/pl.cgi?57563

Bartneck, Ch., Yogeeswaran, K., Ser, Q.M., Woodward, G., Sparrow, R., Wang, S. & Eyssel, F. (2018). Robots and Racism. *Proceedings of 2018 ACM/IEEE International Conference on Human Robot Interaction (HRI’18)*, 1-9. Retrieved from https://doi.org/10.1145/3171221.3171260

Bąk, A. (2013). *Microeconometric methods of researching consumer preferences using the R program.* Warsaw: C.H. Beck publishing.

Bonnemains, V., Saurel, C. & Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology,* 20, 41-58. https://doi.org/10.1007/s10676-018-9444-x

Bostrom, N. & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316-334). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139046855.020

British Standards Institute (2016). *BS 8611:2016. Ethical design and application of robots*.

Brunius, T. (1958). Jeremy Bentham's Moral Calculus. *Acta Sociologica* 3(1), 73-85. https://doi.org/10.1177/000169935800300107

Dennis, L.A., Fisher, M. & Winfield, A.F.T. (2015). *Towards verifiably ethical robot behaviour*. Retrieved from http://arxiv.org/abs/1504.03592

Di Fabio, U., Broy, M. & Brüngger, R.J. (2017, June). Ethics commission automated and connected driving. *Federal Ministry of Transport and Digital Infrastructure of the Federal Republic of Germany*. Retrieved from https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission-automated-and-connected-driving.pdf

Dignum, V. (2017). Responsible autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI'2017)*, 4698-4704. https://doi.org/10.24963/ijcai.2017/655

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology,* 20, 1-3. https://doi.org/10.1007/s10676-018-9450-z

Green, B.P. (2017, November, 3-4). Some Ethical and Theological Reflections on Artificial Intelligence. In *Graduate Theological Union, Pacific Coast Theological Society*, Berkeley. http://doi.org/10.12775/SetF.2018.015

Holbo, J. (2002). Moral Dilemmas and Deontic Logic. *American Philosophical Quarterly,* 39, 259-274.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019). *Ethically aligned design*. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf

Kaplan, A. & Haenlein, M. (2019). Siri, Siri in my Hand, who's the Fairest in the Land? On the Interpretations, Illustrations and Implications of Artificial Intelligence. *Business Horizons,* 62(1), 15-25. https://doi.org/10.1016/j.bushor.2018.08.004

Korea's Ministry of Commerce, Industry and Energy (2012). *South Korean Robot Ethics Charter*. Retrieved from https://akikok012um1.wordpress.com/south-korean-robot-ethics-charter-2012

McCarthy, J.J., Minsky, M.L. & Rochester, N. (1959). Artificial intelligence. *Research Laboratory of Electronics Progress Report No 53.* http://hdl.handle.net/1721.1/52263

McConnell, T. (2018). Moral dilemmas. In: E. N. Zalta (Eds.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/fall2018/entries/moral-dilemmas

Ogryczak, W. (1997). *Multi-criteria linear and discrete optimization: models of preferences and applications to support decisions.* Warsaw: University of Warsaw Press.

Omohundro, S.M. (2008). The Basic AI Drives. In *Proceedings of the 2008 conference on Artificial General Intelligence*. IOS Press, 483-492. https://doi.org/10.5555/1566174.1566226

Project of IEEE Ethics in Action in Autonomous and Intelligent Systems: https://ethicsinaction.ieee.org

Project of IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

Project of Moral Machine: https://www.moralmachine.net

Project of Partnership on AI: https://www.partnershiponai.org

Sinnott-Armstrong, W. (1988). *Moral Dilemmas.* Oxford: Basil Blackwell.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M. & Teller, A. (2016). Artificial Intelligence and Life in 2030. *One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel,* 52.

Thomson, J.J. (1985). The Trolley Problem. *Yale Law Journal,* 94, 1395-1415.

Tørresen, J. (2020, January 7). AI Ethics: How to achieve ethically good artificial intelligence research and development. *AI Ethics Seminars at Chalmers.*

Vamplew, P., Dazeley, R., Foale, C., Firmin, S. & Mummery, J. (2018). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology,* 20, 27-40. https://doi.org/10.1007/s10676-017-9440-6

Veruggio, G. (2006). The EURON Roboethics Roadmap, *6th IEEE-RAS International Conference on Humanoid Robots*, 612-617. https://doi.org/10.1109/ICHR.2006.321337

Warren, M.A. (1997). *Moral Status: Obligations to Persons and Other Living Things. Issues in Biomedical Ethics*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198250401.001.0001

Winfield, P.A. (2017). *ELS issues in robotics and steps to consider them. Part 3: Ethics.* Retrieved from https://www.eu-robotics.net

Wong, P.H. (2016). Responsible innovation for decent non liberal people: a dilemma? *Journal of Responsible Innovation*, 3(2), 154-168. https://doi.org/10.1080/23299460.2016.1216709

Xue, Z., Dong, Q., Fan, X., Jin, Q., Jian, H. & Liu, J. (2017). Fuzzy logic-based model that incorporates personality traits for heterogeneous pedestrians. *Symmetry*, 9(10), 239. https://doi.org/10.3390/sym9100239

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control,* 8, 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X

ARE WE IN THE DIGITAL DARK TIMES? HOW THE PHILOSOPHY OF HANNAH ARENDT CAN ILLUMINATE SOME OF THE ETHICAL DILEMMAS POSED BY MODERN DIGITAL TECHNOLOGIES

Damian Gordon, Anna Becevel

Technological University of Dublin (Ireland)

Damian.X.Gordon@TUDublin.ie; Anna.Becevel@TUDublin.ie

ABSTRACT

Philosophers are not generally credited with being clairvoyant, and yet because they recognise, record and reflect on trends in their society, their observations can often appear prescient. In the field of the ethics of technology, there is, perhaps, no philosopher whose perspective on these issues is worth examining in detail more than that of Hannah Arendt, who can offer real perspective on the challenges we are facing with technologies in the twenty-first century. Arendt, a thinker of Jewish-German origin, student of Martin Heidegger and Karl Jaspers, encountered her life turning point when she was forced into becoming a refugee as the world was shaken by a force of unimaginable brutality that she was one of the first to name "totalitarianism" (Baerh, 2010). She was an independent thinker, separating herself from schools of thought or ideology. Investigating totalitarianism was her ruling passion, and as such her political thought often overshadows her major contribution to other branches of philosophy. Arendt is best known for her accounts of Adolf Eichmann and his trial, and the concept of "banality of evil", though her perspective on politics was driven by a precise and original theory of action. While the latter is inextricably connected to her political perspective, it is also supported by a sharp ontological reflection of social structures and anthropological reflections.

INTRODUCTION

In her 1963 book "*Eichmann in Jerusalem: A Report on the Banality of Evil*" Arendt introduced the notion of the "banality of evil", meaning that very often evil acts are not committed by fanatics or psychopaths, but instead by very normal people who rely on simplistic clichés to justify their actions, rather than thinking for themselves. The concept is not meant to indicate the the deeds of Eichmann and others were in any way ordinary, but that the self-justification they used and complacency of their acts was wholly unexceptional Later in 1978 Arendt warned that "*clichés, stock phrases, adherence to conventional, standardized codes of expression and conduct have the socially recognized function of protecting us against reality, that is, against the claim on our thinking attention that all events and facts make by virtue of their existence*." This, unfortunately, seems to marry well with statements made by ex-employees of social media companies in the past 10 years, and most notably on the 2020 Netflix docudrama "The Social Dilemma" where former employees of social media companies like Facebook, Google, Twitter, Mozilla, and YouTube describe how their companies unthinkingly nurture addiction to their product, and help spread conspiracy theories and disinformation. This film also explores the issue of the impact of social media on user's mental health (and particularly the mental health of adolescents and rising teen suicide rates). The film cites key statistics such as a 62% increase in hospitalizations for American females aged 15–19 and a 189% increase in females aged 10–14 due to self harm, beginning in 2010–2011, which the companies are aware of, but their employees justify

their actions with mindless clichés like "*it's good for my career*", "*I'm just doing my job*", or "*I was only following orders*".

PROPAGANDA

Arendt (1951) said "*In an ever-changing, incomprehensible world the masses had reached the point where they would, at the same time, believe everything and nothing, think that everything was possible and that nothing was true. ... Mass propaganda discovered that its audience was ready at all times to believe the worst, no matter how absurd, and did not particularly object to being deceived because it held every statement to be a lie anyhow. The totalitarian mass leaders based their propaganda on the correct psychological assumption that, under such conditions, one could make people believe the most fantastic statements one day, and trust that if the next day they were given irrefutable proof of their falsehood, they would take refuge in cynicism; instead of deserting the leaders who had lied to them, they would protest that they had known all along that the statement was a lie and would admire the leaders for their superior tactical cleverness*."

Gaber and Fisher (2021) looked at political lying in general, and specifically at the lies told during the referendum over *Brexit* (the proposed withdrawal of the United Kingdom from the European Union and the European Atomic Energy Community), as well as the political campaigns of Donald Trump in 2016 and 2020. They argue that until recently politicians would avoid telling outright lies as they felt the political consequences would be too severe, however in the past decade there has been a change, and politicians have learned to use "strategic lies" (lies that are both attention-grabbing and agenda-setting) with apparently no consequences. They argue that this approach has its roots in the practice of "spin" which grew significantly in the political sphere in the 1990s (Street, 2011), where media advisors presented biased interpretations of events to influence public opinion about a particular issue. Since then, with the growth of social media and increased professionalization of media advisors, "spin" has been transformed into "strategic lying". An early example of this new form of lying was Donald Trump's claim to have "proof" that Barack Obama was not born in the United States (starting the "birther" movement), and he claimed that he was going to send a team of private investigators to Hawaii to explore the truth of these claims, and would donate $5 million to charity if definitive evidence was found that President Obama was, in fact, born on the USA. There is no record of a team of private investigators to Hawaii nor is there no record of Donald Trump donating $5 million to charity in spite of the fact that President Obama did publish his birth certificate, however, Donald Trump continued to reiterate his claims after the evidence, to create the impression of Obama "otherness" in the mainstream media. A similar example occured when current British Prime Minister Boris Johnson, who in his former role as a journalist made false claims about the European Union, including that they were going to ban certain electrical appliances (including vacuum cleaners, kettles, toasters and lawnmowers), and ban bananas that were too bendy. These claims were fact-checked as false a number of times, but nonetheless Johnson continued to reiterate them, including during the runup to the Brexit referendum. Additionally Johnson repeatedly made a claim that the UK sends £350 million a week to the EU, which is a gross figure, in reality the UK sent a net figure of £210 million a week to the EU. Johnson even went so far as to print this claim on the side of a bus, and claim that money could go to the UK national health services instead (See Figure 1 below).

Figure 1. The "Brexit" Bus

As Alastair Campbell (former Press Secretary and Director of Communications to former British Prime Minister, Tony Blair) stated: "*I am afraid we have entered a post-truth, post-shame world. The Washington Post says Donald Trump tells 12 lies a day. His predecessors would have been hounded out of office for one in a term. Boris Johnson won a referendum by lying. His reward? He was made Foreign Secretary and he is now going to be the Prime Minister. There is no shame!*" [21].

Unfortunately this appears to be correct for some people, it appears that for some of the voting public, they have been convinced that we live in a "post-truth" era, where people believe that truth is a relative concept, and they feel empowered to choose their own version of reality, where existing beliefs and prejudices are more important than facts, particularly if the existing beliefs and prejudices are reiterated and amplified by their political leaders. This also means that those individuals have abandoned conventional criteria of evidence and fact-checking, and in exchange they are not obliged to have to think about difficult or unsettling realities (Lewandowsky, et al., 2017). This "filter bubble" is comforting to those who live in them, and the residents of these isolationist spaces tend to reward the politicians who help maintain these bubbles.

SOCIAL MEDIA

Combining two quotes from Arendt, in 1963 she said that "*The trouble with Eichmann was precisely that so many were like him, and that the many were neither perverted nor sadistic, that they were, and still are, terribly and terrifyingly normal. From the viewpoint of our legal institutions and of our moral standards of judgment, this normality was much more terrifying than all the atrocities put together*" and in 1978 she warned "*The sad truth of the matter is that most evil is done by people who never made up their minds to be or do either evil or good*". This could very well be used as a lens for both the developers of social media, and the users of social media. It is important to recognise that social media companies want to keep their users on their sites for as long as possible, therefore they use manipulative approaches, such as "Digital Nudges" (Acquisti 2009) which are small interventions that guide choices without restricting them, such as timely reminders, personalized messages, or small digital rewards. As users are using social media, more and more behavioural data is being collected about them so that an increasingly complex and comprehensive digital model of each individual is created, and the correct means to extent that user's session time will be identified to expose that user to as much advertisement as possible. This, in and of itself, might seem innocuous, but when people are using a range of social media platforms, this can have unintended, catastrophic consequences.

[21] Alistair Campbell, Depression and the politics of mental health: Alastair Campbell on ABC Radio National Breakfast, July 22, 2019

Research by McHugh, et al. (2018) suggests that social media usage can cause symptoms of post-traumatic stress disorder (PTSD) in adolescents, and they also found that these adolescents engage in coping mechanisms to help to reduce the long-term negative effects of exposure.

In the context of lies being spread by social media, research by Vosoughi et al. (2018) indicated that lies spread "*significantly farther, faster, deeper, and more broadly than the truth*". They indicate that there are two key reasons for this: (1) *Confirmation Bias*, people tend to notice and remember things that help confirm their own worldview, and (2) *Repetition*, the more often a lie is repeated, the more likely it is to be believed, even it is refuted each time, because this lie has already been processed by the brain it takes no additional cognitive load to process it again (e.g. the way Donald Trump kept claiming the election was "stolen" in 2020). The individuals creating the lies know that they are doing harm, but the (potentially) millions of people who believe the misinformation, who spread this misinformation, and who embroider the misinformation, are not intending to do harm, but are part of a larger process that is detrimental to all participants.

MACHINE LEARNING

Famously Arendt (1962) wrote "*I have always believed that, no matter how abstract our theories may sound or how consistent our arguments may appear, there are incidents and stories behind them which, at least for ourselves, contain as in a nutshell the full meaning of whatever we have to say. Thought itself - to the extent that it is more than a technical, logical operation which electronic machines may be better equipped to perform than the human brain arises out of the actuality of incidents, and incidents of living experience must remain its guideposts by which it takes its bearings if it is not to lose itself in the heights to which thinking soars, or in the depths to which it must descend.*". This is a profound insight into the problems of machine learning, Arendt is arguing that real thinking can only occur through the lens of human experience, and an abstract representation of ideas do not in fact encompass the totality of thinking. The world Arendt describes is a vively and turmoiled one, where each individual acts freely in their environment while simultaneously creating a shared political space, a world that our current technologies seem unable to describe at this stage due to the limitations of machine learning. The concept of Machine Learning was developed by Samuel (1959), and generally consists of the following steps (Langley, 2011):

1. Collecting data about a significant number of examples of particular scenario; the data usually consists of key descriptors or characteristics;
2. The data is analysed using a computer program that attempts to uncover rules or relationships between the descriptors;
3. The rules are then used to predict the outcomes of new examples of the scenario that haven't been presented to the computer program yet.

This approach has led to a growing catalogue of disastrously poor results, for example, in 2014 Amazon began developing a computer program to help in personnel recruitment, and after a year they discovered that the system was sexist in operation, and would always prefer male candidates to female ones, and eventually they abandoned that system. What a subsequent analysis found was that because a significant majority of existing successful candidates were male, the system was fed an abundance of data on male candidates and less on female candidates (Fumiko, et al., 2020). In 2013 IBM partnered with *The University of Texas MD Anderson Cancer Center* to develop a new "Oncology Expert Advisor" system that would ultimately lead to a cure for cancer. Unfortunately, the resulting system gave

erroneous, and downright dangerous cancer treatment advice, and had to be finally abandoned in 2018, simply because the IBM engineers trained their software on synthesized data, rather than real patient data (Strickland, 2019). Hendrycks, et al. (2019) set out to show the limitations of machine learning algorithms, by selecting 7,500 specifically curated images of a large dataset of images of animals, insects and other natural phenomena, they reduced the effectiveness of a machine learning algorithm from 92% to 2%.

The problem with these systems is that they rely almost completely on data to draw their conclusions, and if data is misconfigured, then the rules that the system deduces are flawed. Additionally, it is only possible for some machine learning systems to express the rules that they have deduced in a manner that a human being can understand, for other systems the manner in which they deduce and encode these rules cannot be expressed as text, and they are therefore said to lack *explainability* (London, 2019). This is a very serious issue, if the systems can't even explain why they are making decisions, it makes trusting those decisions more difficult, so much so that the European Union is regulating the use of machine learning, and requiring that it must be of the explainable variety (Hamon, et al., 2020). As well as bias in data, other issues that appear to cause poor decision-making includes:

- *Underfitting*, is where the rules that the systems deduced aren't a sufficiently detailed model of the complexity of the data presented to the system.
- *Overfitting*, is where the rules that the system deduced are too specifically tailored for the data presented to the system, and can't accurately generalise the lessons learned.
- *Undersampling*, is where the distribution of data in one characteristic of the dataset doesn't reflect the population under investigation because one group is under-sampled, for example, if one race of people is under-represented in a dataset about a group of people.
- *Oversampling*, is where the distribution of data in one characteristic of the dataset doesn't reflect the population under investigation because one group is over-sampled, for example, if one race of people is over-represented in a dataset about a group of people.
- *Proxy Variables*, is where you have to use a stand-in variable because it isn't possible to represent a characteristic directly. So, for example, if you can't measure people's level of health, it might be easier to measure how much money people spend on health, as a proxy to measure level of health. Unfortunately, this doesn't take into account wealth level.
- *Missing Variables*, is where the characteristics selected in the dataset are not everything that should be taken into account to have a representation sample.
- *Underspecification*, identified in 2020, is where the characteristics chosen in the dataset don't represent the totality of the key features required to model the data (D'Amour, et al., 2020).
- *Data Scarcity*, is where insufficient data is presented to the system, and therefore, there isn't enough variation in the data to represent all of the potential cases the system will encounter.

These all simply point to an inherent flaw in the development of machine learning systems, that unless the exact parameters of the problem are already fully understood, it might not be possible to identify the correct dataset characteristics to accurately represent the problem. The truth of the situation is that the datasets used by these systems cannot capture the full diversity of real-world experience. When considering the phenomenological nature of action (Dal Lago, 2016), not being able to describe the complexity of human experience doesn't only mean missing on diversity, but missing on the chance to obtain it at any stage. Human experience is the way through which agents reveal themselves and

simultaneously accept the risks implied by this revelation. The exposure of human experience is a necessary and sufficient condition to create a political space where the individuals can work-together and regulate themselves in environments not regulated by governments such as the internet. (Arendt, 1958). Moreover, the people who create and curate datasets bring with them a series of tacit assumptions, and even cognitive biases, about the problem that make a representative dataset less possible. One common erroneous assumption that many people make is how frequently unusual events occur (Paulos, 1988), and this can led to the creation of unrepresentative datasets; again as Arendt says: "*incidents of living experience must remain its guideposts by which it takes its bearings*".

Unfortunately, modern technology is contributing to cognitive biases, for example, since 2009 the Google search engine has incorporated a "Personalized Search" which means that results returned are not the same for everyone, instead they are based on each individual user's personal behaviour and interests as well as those of the user's social circle (Zamir and Korn, 2020). This creates a "filter bubble" that creates polarization and echo chambers, and results in an exogenous isolation effect, as well as a lack of full discussion of the topics (Min, et al., 2019). This issue was highlighted by Arendt's when she stated that: "*To hold different opinions and to be aware that other people think differently on the same issue shields us from Godlike certainty which stops all discussion and reduces social relationships to an ant heap*".

CONCLUSIONS

These issues are a small sampling of the perspective and insight that Arendt can give us on computer ethics, and her reflections can be both thought-provoking and illuminating in terms of how we should develop and use new technologies. As mentioned at the start, philosophers are not generally credited with being clairvoyant, and yet Arendt's perspectives might provide a way forward in the modern world. And her work, and the work of other 20th century philosophers, urgently need to be re-examined in the light of the serious political decisions that are being made by so many in such a mindless way.

KEYWORDS: Digital Ethics, Hannah Arendt, Machine Learning, Conglomerations.

REFERENCES

Acquisti, A. (2009) "Nudging privacy: The behavioral economics of personal information", *IEEE Security & Privacy*, 7(6), 82-85.

Arendt, H. (1951). *The Origins of Totalitarianism*. New York: Schocken.

Arendt, H. (1958). *The Human Condition*. Chicago: University Chicago Press.

Arendt, H. (1962). Action and the Pursuit of Happiness. In A. Dempft, H. Arendt & F. Engel-Janosi (Eds.), *Politische Ordnung und Menschliche Existenz. Festgabe für Eric Voegelin zum 60 Geburtstag* (pp. 1–16). Munich: C. H. Beck.

Arendt, H. (1963). *Eichmann in Jerusalem*. New York: Viking Press.

Arendt, H. (1978). *The Life of the Mind*. Harcourt Brace Jovanovich.

Baerh, P. (2010). *Hanna Arendt, Totalitarianism, and the Social Sciences*. Stanford: Stanford University Press.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M.D. & Hormozdiari, F. (2020). Underspecification presents challenges for credibility in modern machine learning. Retrieved from https://arxiv.org/abs/2011.03395v2

Dal Lago, A. (2016). [Introduction]. In H. Arendt, *Vita Activa: la condizione umana* (pp. 6-38). Milano: Bompiani.

Farhall, K., Carson, A., Wright, S., Gibbons, A., Lukamto, W. (2019) "Political Elites' Use of Fake News Discourse Across Communications Platforms", *International Journal of Communication*, 13, 23.

Fumiko, K., Arai, H. & Ema, A. (2020). Ethical Issues Regarding the Use of AI Profiling Services for Recruiting: The Japanese Rikunabi Data Scandal. Retrieved from https://arxiv.org/abs/2005.08663

Gaber, I., Fisher, C. (2021). ""Strategic Lying": The Case of Brexit and the 2019 UK Election", *The International Journal of Press/Politics*, 1940161221994100.

Hamon, R., Junklewitz, H. & Sanchez Martin, J. (2020). Robustness and Explainability of Artificial Intelligence. EUR 30040 EN, JRC119336. Luxembourg: Publications Office of the European Union. Retrieved from https://publications.jrc.ec.europa.eu/repository/handle/JRC119336

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2019). Natural adversarial examples. Retrieved from https://arxiv.org/abs/1907.07174

Langley, P. (2011). The changing science of machine learning. Machine Learning, 82(3), 275–279.

Lewandowsky, S., Ecker, U. K., Cook, J. (2017) "Beyond misinformation: Understanding and coping with the "post-truth" era". *Journal of Applied Research in Memory and Cognition*, 6(4), pp. 353-369.

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, 49(1), 15-21.

McHugh, B. C., Wisniewski, P., Rosson, M. B., Carroll, J. M. (2018) "When social media traumatizes teens: The roles of online risk exposure, coping, and post-traumatic stress", *Internet Research*, 28(5), pp. 1169-1188.

Min, Y., Jiang, T., Jin, C., Li, Q., & Jin, X. (2019). Endogenetic Structure of Filter Bubble in Social Networks. *Royal Society Open Science*, 6(11), 190868.

Paulos, J. A. (1988). *Innumeracy: Mathematical Illiteracy and its Consequences.* New York, NY: Hill and Wang.

Samuel, A. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3): 210–229.

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, 56(4): 24-31.

Street, J. (2011) *Mass Media, Politics and Democracy*, Palgrave Macmillan.

Vosoughi, S., Roy, D., Aral, S. (2018) "The Spread of True and False News Online", *Science,* 359(6380), pp. 1146–51

Zamir, O., & Korn, J. (2020). *U.S. Patent No. 10,691,76*5. Washington, DC: U.S. Patent and Trademark Office.

CHECK YOUR TECH - WHOSE RESPONSIBILITY IS IT WHEN CYBERHARASSMENT OCCURS?

Dympna O'Sullivan, Damian Gordon, Michael Collins, Emma Murphy

Technological University of Dublin (Ireland)

Dympna.OSullivan@TUDublin.ie, Damian.X.Gordon@TUDublin.ie,
Micheal.Collins@TUDublin.ie, Emma.X.Murphy@TUDublin.ie

ABSTRACT

Social media has become a dominant aspect of many people's lives in many countries. Unfortunately that resulted in widespread issues of bullying and harassment. While frequently this harrassment is intentional, there have been occasions where automated processes have been inadvertently responsible for this sort of harassment. The software tools that allow people to harass others could have further features added to them to reduce the amount of harassment that occurs, but more often than not, where programmers are developing these systems then don't anticipate the range of ways that these technologies will be used (this is called "consequence scanning"). The authors of this paper are developing a new digital ethics curriculum for the instruction of computer science students. In this paper we present two case studies we have developed with a focus on cyberharrassment. Each case study is accompanied by a list of specific questions to be used by the instructor to allow students to evaluate the implications of developing social media systems as well as a generic case studies checksheet that allow deeper reflection on the intended and unintended consequences of introducing new technologies.

INTRODUCTION

Cyberharassment has grown enormously as the online world continues to grow on an annual basis (Smith, et al., 2008). The impacts of this form of harassment can be extremely serious on its victims, including issues such as anger, frustration, depression, low self-esteem and suicidal ideation (Hinduja and Patchin, 2009). The situation has become so serious that a number of national and international organisations have been founded in the past decade to help combat this issue, and to raise awareness of its effects, including the Cybersmile foundation[22], the Online Abuse Prevention Initiative[23], and the Cyber Civil Rights Initiative[24].

Legislation has been introduced in different countries to help ameliorate the impacts of cyberharassment, including the *Philippines' Cybercrime Prevention Act of 2012*, and the *Protecting Canadians from Online Crime Act (2014)*. In December 2020, Ireland signed into law the *Harassment, Harmful Communications and Related Offences Bill*, which provides the first legal definition of cyberharassment in Irish law, and includes penalties of up to 10 years incarceration for people who engage in egregious behaviour, particularly, so-called "revenge porn". Additionally, social media companies have added features to their systems to help combat harassment, and they typically use a combination of artificial intelligence (AI) and professional moderators to review and remove

[22] https://www.cybersmile.org/

[23] https://onlineabuseprevention.wordpress.com/

[24] https://www.cybercivilrights.org/

inappropriate content. Unfortunately, there are issues with the moderation process; the scale of the task is enormous, and the moderators are often hired based on the lowest salary, and may lack knowledge of the platform-specific guidelines, as well as the linguistic fluency in the language of the content (Roberts, 2019).

Women are disproportionately affected by cyberharassment, in fact, the United Nations Broadband Commission Working Group on Gender indicated that 73% of women worldwide have been exposed to or have experienced some form of online violence (UN Broadband Commission for Digital Development, 2015). The WWW Foundation has found that law enforcement agencies and the courts are failing to take appropriate actions for cyberharassment against women in 74% of 86 countries surveyed (World Wide Web Foundation, 2015). The sheer volume of cyberharassment experienced by women has significant social and economic implications for women's status on the Internet. These include time, emotional bandwidth, financial resources including legal fees, online protection services, and missed wages. This is a problem that needs to be addressed if social media is to remain an open and empowering space for women and girls, and by extension, for boys and men.

This issue is one of grave concern, and is one of a rapidly growing number of computer ethics issues that have been emerging recently, to such an extent that a number of third-level institutes across Europe are collaborating to explore some of these key ethical challenges, and to develop educational content that is both based on pedagogically sound principles, and motivated by international exemplars of best practice to highlight these matters as part of the Erasmus+ Ethics4EU project[25] (O'Sullivan and Gordon, 2020). One specific development that is being undertaken is the creation of a lesson focusing on social media, and concentrating specifically on the ethics of developing social media software that can have a negative impact on people's lives.

Part of the lesson is the development of, specifically synthesized or fictionalized case studies, that focus on different types of cyberharrassment. These are to help computer science students at consequence scanning – a way to consider the potential consequences - intended and unintended - of a new technological product or service on people, communities and the planet (Doteveryone, 2020). The case studies are suitable because they provide a way to examine a specific phenomena with a focus on interpreting events, and exploring the societal context in which the case occurs (Martin, et al. (2018). Also because these cases are qualitative, they will be somewhat novel in computer science courses which are typically more quantitative in nature. They can be used to both explore and evaluate specific problems and challenges of social media tools, as well as exploring digital ethics in a more general context.

The two case studies we have developed are a part of a wider curriculum on digital ethics for computer science students. The case studies concern the impact of technology on people's lives, and how it can adversely impact people's lives both deliberately and accidentally. Each case study comprises a detailed narrative and set of questions (or "Talking Points") to be used by an instructor in delivering the content. We have also developed a generic case studies checksheet that allows a student to examine any scenario using a range of criteria that explore the features of the case and the consequences of the technology - intended and unintended. The checksheet is based on work by Yin (2017) and is intended for deeper reflection on specific aspects of the case studies and is to be used in conjunction with the "Talking Points" outlined above.

The first case study focuses on a deliberate harassment scenario, where one individual sets out to harm another person using a range of on-line tools, including social media systems, and is presented in the form of an epistolary, in this case a collection of emails. The second case study explores how a

[25] http://ethics4eu.eu/

combination of minor technical issues can result in catastrophic consequences for a family, and this is presented in the style of a newspaper article.

As mentioned above, this content is designed for computer science students to allow them explore how actions in the online world can have calamitous consequences for people in the real world. This is very important, particularly for computer science students who could potentially do the most damage if they chose to engage in harassment (for example, using photo manipulation software, and deepfakes (see Tolosana et al., 2020)) but more importantly, they must have an awareness of these issues since they are going to be the people building the next generation of software systems and social media tools.

METHODS

A case study is a suitable vehicle for examining this topic as case studies explore specific real-world phenomena that focus on interpreting events, and exploring the societal context in which the case occurs (Martin et al., 2018). The qualitative nature of these cases can be seen as novel when introduced in computer science courses which are typically more quantitative in nature. They can be used to both explore and evaluate specific problems and challenges of introducing new technologies into developing countries, as well as exploring digital ethics in a more general context. The materials allow for detailed examination of technological, organizational and social implications.

In this section we introduce two case studies we have developed as a part of a wider curriculum on digital ethics for computer science students. The case studies concern the impact of cyberharrassment on people. Each case study comprises a detailed narrative and set of questions (or “Talking Points”) to be used by an instructor in delivering the content. We also introduce a general checksheet that can be used by students to evaluate scenarios involving the development of new digital products and services.

These case studies have been developed specifically as teaching tools; each is based on a synthesis of several real cases, and are designed to generate detailed and diverse discussions by student groups about the ethics around these scenarios. The use of synthesized case studies has a long history in the teaching, particularly in Law courses (Dyer et al., 1997) as they can help to avoid issues such as confidentiality and legal privilege, which are clearly very important considerations when discussing in this particular context, that of cyberharassment. To highlight the fictitious nature of the case studies, pre-existing fictional characters and place names are used to underscore the fact that these case studies are not real.

These synthesized case studies somewhat resemble a teaching approach that is already used in computer science, the “toy problem”, which is an approach used in the teaching of computer programming, where a scenario is created as an expository device to help students explore challenges around a specific programming problem (Pearl, 1984). These problems often distil some key features or challenges into simplified scenarios, and sometimes combine several distilled features into one problem that would be unlikely to occur in a real-world setting but they are very useful in teaching students about the challenges in that specific domain. Thus, these case studies are designed in the same way to highlight specific features or challenges that serve as the basis for the talking points to discuss ethical topics with the students.

The first case study focuses on how technology can be used to deliberately harass and intimidate people, whereas the second case study explores how technology can devastate people’s lives.

CASE STUDIES

Case Study 1

The first case study concerns the deliberate harassment of one individual by another, and the full case study, told in the form of an epistolary (with emails, receipts and other records), can be found here: http://damiantgordon.com/Ethics4EU/Cyberharassment/CaseStudy1.pdf

Case Study 1 Summary

- Lucy Honeychurch and George Emerson had been dating but they have broken up. Lucy wants them to take a break from communicating for a few months but George is going to do everything he can to get them to talk again.
- George starts by topping up Lucy's credit for her recycling company, which she thanks him for but reiterates her desire for them to take a break from communicating. George makes a fake apology and writes a positive review about her on her company's website, expecting her to thank him for it.
- After a week George sends an email demanding an acknowledgement of his review and Lucy again reiterates her desire for them to take a break from communicating.
- George takes a week off, but then tries to get Lucy to talk to him again by trying to give her a late birthday gift. She tells him not to give it to her and to stop bothering her and her friends, reminding him that his controlling behaviour is the reason they broke up in the first place.
- George reacts badly and begins to take a more aggressive stance and falls in with an incel group who encourage his bad behaviour. He escalates his stance with her, harassing her online and in the real world until eventually his cyberharassment results in her reporting him to the police, resulting in his eventual arrest.

Case Study 1 Talking Points

1. Should the police have more powers in terms of being able to intervene in cyberharassment situations, including the ability to seize devices that are suspected of being used in these cases? Why?
2. Should the government pass laws that would result in bigger sentences for people who engage in cyberharassment as a deterrent? Why?
3. Do you think that social media companies have a great obligation to protect their users, for example, should they allow users to block specific IP addresses or phone numbers? Why?
4. Should people who engage in cyberharassment be banned from certain types of jobs, for example, law enforcement or the civil service? Why?
5. Should people who engage in cyberharassment be banned from some social media sites? Why?

Case Study 2

The second case study concerns the accidental harassment of a family by a group of technology companies, and the full case study, told in the form of a newspaper article, can be found here: http://damiantgordon.com/Ethics4EU/Cyberharassment/CaseStudy2.pdf

Case Study 2 Summary

- The Harris family consists of parents Billy and Donna and their two sons, Buck and Harry. Due to a glitch in location-mapping software, their address is incorrectly given for food deliveries, law enforcement issues, credit card applications and for all manner of other deliveries and computer issues.
- Donna and Harry are trying to do something about the situation and deal with the fallout of this "minor technical error" (including several lawsuits), whereas Billy and Buck are burying their heads in the sand about the situation.
- A newspaper reporter, Hildy Johnson, comes to stay with the family to document their situation but becomes so outraged that she commits the resources of her newspaper's legal team to help them, as well as a friend of hers who is a hacker.
- Hildy's presence in the family gives Donna the confidence to kick Billy out, who moves into an apartment with Buck.
- With the help of Hildy's newspaper, some of the computer companies responsible for a lot of the problems that the Harris family have been experiencing give Donna $53 million, with no admission of liability. Donna and Harry move to Beverly Hills and the mapping company (TendreMaps) who were the principal culprit have their systems hacked that changed the "minor technical error" to send all the wrong deliveries to the headquarters of TendreMaps instead.

Case Study 2 Talking Points

1. Do the programmers who wrote the software that set the default IP address to 0.0.0.0 bear responsibility to what happened to the Harris family? Why?
2. Does the TendreMaps Mapping Company bear responsibility to what happened to the Harris family? Why?
3. When the Harris family got internet access in their house they signed a contract with Terms & Conditions that clearly state that there is no liability for any problems caused by errors in software. Even if such agreements are legal, are they ethical? Why?
4. Hildy suggests that computer companies pay their computer programmers poorly but at the same time spend millions on legal teams to defend themselves. Is this a good business model? Why?
5. Some people have suggested that Harry worked with Henry Dorsett Case to redirect the mapping from the Harris House to the TendreMaps headquarters. If so, do you think it was justified? Why?

Case Study Checksheet

A task sheet for students to work through several times and internalise

Evaluation criteria	Notes
What is the case study about?	Introduction:
What is the organisation?	Introduction:
What are the technology issues?	Introduction:
Who are the principal actors?	Introduction:
What types of data were collected?	Data Collection:
From which sources did they come?	Data Collection:
How was the data recorded?	Data Collection:
What was the situation previously?	Main Features:
What interventions have been introduced?	Main Features:
What were the general outcomes of this intervention?	Main Features:
Are there any legal, social or ethical issues associated with this intervention?	Main Features:
Is there a chronological or other logic sequence for analysis?	Main Features:
What is the nature of the organisation?	Organisation:
What is its history?	Organisation:
How is it structured?	Organisation:
How has it changed as a result of intervention?	Organisation:
Who are the principal actors in detail?	People (Ecology):
What are their positions within the organisation?	People (Ecology):
What are their technical skills?	People (Ecology):
Does the target population for this intervention include more people?	People (Ecology):
What technology was present? What software? What hardware?	Technology:
What technical level of expertise exists within the organization?	Technology:
What new technology has been introduced for this intervention?	Technology:
How has the new technology effected the organisation?	Technology:
What are the possible consequences of this technology - intended and unintended?	Technology:
How successful has the intervention been?	Evaluation:
What new outcomes have been identified?	Evaluation:
What went well in this intervention?	Evaluation:
What did not go well in the intervention?	Evaluation:
What alternative approaches could have been taken?	Evaluation:

DISCUSSION

Cyberharassment is an extremely serious issue and is something that software developers need to take into consideration when they are creating new systems that allow users to interact. As mentioned previously, the impact of harassment includes issues such as anger, frustration, depression, low self-esteem and suicidal ideation. It is therefore incumbent that software developers reflect seriously on the ways their creations will be used.

We have presented two case studies, a set of case study specific questions (“Talking Points”) and a generic case study checksheet to be used in the instruction of computer science students to allow them to reflect on the consequences - intended and unintended - of new technologies for use in developing world contexts. Although the content is developed for computer science students, it could be adapted for other educational disciplines. All of the synthesized case studies that are being designed as part of the Ethics4EU project are created in pairs. The first is usually more straightforward, focusing on the more traditional perspective on cyberharrassment (one person harassing another), whereas

the second one is looking at the accidental harassment of a family (in this case, an IP address issue). In this way, they work well as individual case studies, but also when taken as a pair they provide an interesting contrast.

After piloting these case studies with a small classgroup, some benefits of the synthesized case studies became clear; students commented that because they knew the scenarios were fictitious, they felt more comfortable elaborating new details about the cases and they also felt more comfortable hypothesizing motivations of particular actors in the scenarios. They also commented that the case studies opened their eyes to some of the problems associated with technology that they had not thought of before. They highlighted the notion that the first case study was the result of a single person's deliberate actions (and therefore, the individual responsible is clear), but in the second case study, it was as a result of the accidental side-effects of a group of organisations' technical decisions (and therefore, the responsibility is distributed, and unclear). A few commented that the use of pre-existing fictional place names made them curious to follow-up on those references, and to explore some literature.

In future work, we intend to develop a larger range of educational content for the instruction of digital ethics. Content will focus on pertinent issues such as privacy, computer security, surveillance and facial recognition, the Internet of Things, AI and algorithmic decision-making including biases such as racial and gender biases often present in large datasets and the environmental implications (specifically the carbon footprint) of storing excessive quantities of data in data centres. We intend to evaluate the educational materials with students in the classroom, gathering feedback from students on the educational instruments and evaluating their before-and-after understanding of the ethical issues raised in the case studies.

ACKNOWLEDGEMENT

The authors of this paper and the participants of the Ethics4EU project gratefully acknowledge the support of the Erasmus+ programme of the European Union. The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

KEYWORDS: Digital Ethics, Cyberharassment, Accidental harassment, Consequence scanning.

REFERENCES

Doteveryone, Consequence Scanning, An Agile Practice for Responsible Innovators https://www.doteveryone.org.uk/project/consequence-scanning/, last accessed 23/12/2020.

Dyer, B., Hughson, M.A., Duns, J., Ricketson, S. (1997) "Teaching Note: Creating a Corporations Law Case Study", *Legal Education Review*, 8

Hinduja, S.; Patchin, J. W. (2009). *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Thousand Oaks, CA: Corwin Press. ISBN 978-1-4129-6689-4.

Martin D.A., Conlon E., Bowe B. (2018) "A Constructivist Approach to the use of Case Studies in teaching Engineering Ethics". In: Auer M., Guralnick D., Simonics I. (eds) "Teaching and Learning in a Digital Worl", ICL 2018. *Advances in Intelligent Systems and Computing*, vol 715. Springer

O'Sullivan, D., Gordon, D. (2020) "Check Your Tech – Considering the Provenance of Data Used to Build Digital Products and Services: Case Studies and an Ethical CheckSheet", IFIP WG 9.4 European Conference on the Social Implications of Computers in Developing Countries, 10th–11th June 2020, Salford, UK.

Pearl, J. (1984) "Heuristics: Intelligent Search Strategies for Computer Problem Solving, Addision Wesley.

Roberts, S.T. (2019) *Behind The Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Smith, P.K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., Tippett, N. (2008) "Cyberbullying: its nature and impact in secondary school pupils". *The Journal of Child Psychology and Psychiatry*. 49 (4): 376–385. doi:10.1111/j.1469-7610.2007.01846.x. PMID 18363945.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. and Ortega-Garcia, J. (2020) "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection", *Information Fusion*, 64, pp.131-148.

UN Broadband Commission for Digital Development, 2015, "Cyber Violence against Women and Girls", Available at: https://www.itu.int/pub/S-POL-BROADBAND.14

World Wide Web Foundation, 2015, "Women's Rights Online - Translating Access into Empowerment", Available at: http://webfoundation.org/docs/2015/10/womens-rights-online21102015.pdf

Yin, R.K. (2017) *Case Study Research and Applications: Design and Methods*, Sage Publications.

ETHICS, INTELLECTUAL CAPITAL AND INTELLIGENT COMPANIES

Carlos Fernández-García, Mario Arias-Oliva, Rubén Fernández-Ortiz, Jorge de Andrés-Sánchez

University Rovira i Virgili (Spain), Complutense University of Madrid (Spain),
University of La Rioja (Spain), University Rovira i Virgili (Spain)

carlos@cruam.com; mario.arias@ucm.es; ruben.fernandez@unirioja.es; jorge.deandres@urv.cat

INTRODUCTION

Today, intellectual capital (IC) is considered the knowledge and experience that a company has, giving its operation the generation of value. In such a way, that the IC is based on the contributions made by its human talent in organizational productivity linked to the relationships that the company maintains with its commercial context. From this perspective, IC relates the knowledge and experience of people with the technology, productivity and competitiveness of an organization, that is, it is an internal and external relationship of the company with the people involved (employees, customers, suppliers, among others) that generates value through knowledge and experience (Azofra, Ochoa, Prieto and Santidrián, 2017; Gómez, Londoño and Mora, 2020).

Due to such appreciations, the existence of IC in companies is considered permanent because it is articulated with productive activities since its creation and it is considered an intangible asset because it adds value to organizational operations by optimizing production processes and business relationships. Under these arguments, the use of CI leads to the strengthening of the company in its position in the competitive market and its financial profitability, in other words, it points to the intangible asset of the company because it means the existence of knowledge attributable to all people belonging to the organization being its strategic nature.

INTELLECTUAL CAPITAL AND KOWLEDGE MANAGEMENT

Articulated with the above, the relevance of IC in intelligent organizations is given by the factors of knowledge and innovation that add value to the entire organizational structure, the performance of human talent and the competitive environment, therefore, IC is identified as a added value element that contributes to knowledge management (KM) as an asset of the company. In accordance with this, CG provides the intellectual capacity of the organization through human talent, involving different sources of knowledge in the business environment, for this reason, people contribute various abilities, skills and knowledge valued as the imperceptible capital that is becomes the relevant element that raises organizational capacity, dividend producers and strategic positioning (Docasal, 2016, Farah y Abouzeid, 2017).

In this regard, IC is identified as a path for QA and organizational development because it explores and uses the most important value: human talent and the knowledge that talent possesses and contributes to the organization. Therefore, IC is made up of consciously integrated structures, methods, and interactions, introduced to produce value to an organization's body of knowledge and experiences. So, promoting the increase of IQ in a company means strengthening its corporate and managerial functioning centered on the individuals who work and interact with the internal and external business environment. From these considerations, the IC is composed of: Human Capital (CH), Relational Capital (CR) and Structural Capital (CE) (Gómez, Londoño and Mora, 2020).

In relation to CH, it can be pointed out that it is constituted by the knowledge and experience of the working personnel within an organization, perceived in the performance of their functions in the workplace, showing: competences, abilities, skills and work capacities to achieve the Business success. Under this argument, the CH is considered particular being in the capacities and knowledge of the worker, differentiating one from another in the action and performance within the organization. Therefore, CH is characterized by knowledge (technical, academic and experience), skills (know-how), attitudes (disposition, effort and behaviors) and values (ethics, responsibility, collaborative work, among others) that people have and that are managed as an advantageous knowledge of the organization (Villegas, Hernández and Salazar, 2017).

From this position, human talent has its own bio-psycho-social characteristics that are manifested in the search for well-being and self-realization, in harmony with the personal potentialities of the current environment: economic, political, environmental, social and legal. Approaching CH in its entirety means transcending the traditional dualism and beginning to see the company from the contributions of training, training and permanent and continuous development of its staff. Under this approach, CH management in an organization focuses its main purpose on developing a workforce of committed, responsible, qualified, motivated and aligned with organizational objectives. In this way, the CH for its internal promotion, considers: results, efficiency and proactive attitude in its time of permanence within this company.

The aforementioned position allows orienting the achievement of the mission, vision and organizational objectives based on the culture and development of proactive human talent, modifying procedures and techniques to adapt to the business environment. In other words, the CH links the company's capacity for results with the aspirations and perceptions of the individual (their objectives and goals, possibilities for growth and group construction). For this reason, in CH, the worker develops the ability to lead action in a specific direction, promoting action values and foreseeing scenarios for improvement. His behavior shows a transforming, competitive, committed, entrepreneurial being and with values within a changing social context and with serious tensions and uncertainties.

In this sense, the CH involves employees in the demands of responsibility and high performance in work relationships, being able to guide the achievement of the mission, vision and organizational objectives based on and the surrounding culture, of the dialogic communication and the development of proactive human talent, modifying procedures and techniques to adapt to current social realities. Their communicational, functional and self-referential relationships allow the recursion of their basic operations and social cohesion (Chiavenato, 2015; Contreras and Rodríguez, 2018).

Now, in intelligent societies and organizations, CH is associated with the innovation and technology of the organization, because the information that is developed is the product of the knowledge and experience of human talent, therefore, the inventiveness of the company it is related to the skill, practice and creativity of the workers. For this reason, innovation and its relationship with CH is given by: 1. The digitization of companies, 2. Globalization and mobility, 3. Aging of the workforce, 4. New job markets and 5. New models of deal. Hence, it follows that the strategic improvement of organizations, CH and their adaptation through ICTs respond to the forefront of the virtual world and the new needs of the competitive market, influencing labor relations and the flexibility of employment with the least social and environmental impact.

In correspondence with Relational Capital (CR), it is based on the position that companies do not have isolated systems, meaning the exchange of knowledge between individuals within the company together with the external relationship. In this sense, the CR adds the assessment of the social interactions of the organization with all the stakeholders (clients, suppliers, workers and shareholders).

Under these arguments, the CR constitutes the valuation of customers when they carry out commercial transactions with their suppliers (Jiménez, 2018; and León, 2020).

From this perspective, the CR is made up of a system of relationships and knowledge that add value to the organization, which are incorporated into the company as a result of its own performance in front of market agents and society. For this reason, the CR links the organization with the knowledge raised in the analogy with: customers, suppliers, shareholders or partners, competition, other companies, entities and government agents and with the corporate image of the organization. In such a way, that the knowledge and information produced are implicit in the relationship system (Carrillo, Bensusán, & Micheli, 2016).

In CR, the knowledge product of the fabric of relationships that the company generates with other organizations (public and private) and people (clients, entrepreneurs, others) is valued through its operations and daily activities. The management of this knowledge is oriented towards the benefit of the company, reflected in market strategies, efficiency in business development, technological and innovative environments, exchange dynamics, among others.

Regarding Structural Capital (CE), organizations have information systems, work procedures, management systems, research and development (R&D) for an effective performance of their operations, therefore, these assets are part of the company and they are legally protected (intellectual property, copyright, among others).

Indeed, the CE represents the knowledge developed in the work routine contributed by human talent and collaborators, therefore, it is linked to the facilities, processes, business policies and technological innovation. Seen in this way, the CE is interpreted as the knowledge transmitted in the internal processes for the continuous improvement of the company using technology as a support tool for its operational activities. This capital incorporates into the organization aspects such as: innovation of products and services, organizational culture, management systems, ICT, intrinsic collaboration and the improvement of productive, functional and operational processes.

For this reason, CE is part of the set of intangible values linked to the agreed methods, acquisition, organization and transfer of knowledge. It is for these appreciations that the knowledge acquired in the intelligent organization is an experiential teaching and a learning instrument. This means that a company that aspires to endure in the current reality must be competent to decipher the demands of the environment and to anticipate them. Hence, the relevance of IC and its three components within a company (Carrillo, Bensusán, & Micheli, 2016).

Now, the organizational culture also belongs to the CE because it depends on the values and norms that intervene in the transmission of knowledge and work interrelationships. This point of view secures the credit of policies and institutes the practice of innovation and productivity. Therefore, the management of ethical values in the CE tends to base decision-making including communication channels (Alarcón, Álvarez, Goyes & Pérez, 2012).

In this argument, an individual who learns in an organization requires ethical values that identify him with the organizational culture. Certainly, the ethics in human talent produces the empowerment of work, strengthens collective habits and influences the interior and exterior of the organization by developing various relationships with the socio-productive context. The relevance of ethics in organizations is given by the attitudes that people assume towards their functions, relationships and responsibilities, influencing the productive activity and the strategic management of the organization (Gómez, Londoño and Mora, 2020).

INTELLECTUAL CAPITAL, ETHICS AND SMART SOCIETIES

So, that relationship between ethics and IQ in the organization transcends social and cultural elements because morality, ethics and values are also part of the intellectual formation of the people who contribute to the development of the company. IC from the intangible asset approach is built based on knowledge and training, based on ethical values for progress in the organization connecting with the CR, CH and CE (Axtle and Acosta, 2017).

Articulated with the above, the analogy of ethics and IC are elements that create value for the individual (endogenous) and the organization (exogenous), reflecting themselves in the productive processes and in the business results. The increase in ethical values and IQ favors the transfer of knowledge simultaneously with the combination of technological resources, emerging an interactive process of exchange between the company and people for the production and commercialization of products and services. This knowledge created within the company expands towards external relations, enhancing its competitive value and leadership position (Jiménez, 2018; and León, 2020).

In such a way that the management of ethics and IC has the purpose of guaranteeing ethical behavior, the development and innovation of processes and the correct decision-making to achieve business objectives in complex productivity. From this point of view, the organizational culture influences the behavior of people, positively affecting the productive system, corporate image and learning capacity, considering an analogy of capacities between people, the organization and society, generating opportunities to learn (intelligence) and act (practice and experience) using Information and Communication Technologies (ICT) (Carrillo, Bensusán, & Micheli, 2016).

Due to such appreciations, the management of ethics and IQ in intelligent people, organizations and societies is developed in the continuous strategies and actions based on the generation and transfer of knowledge. It is at this point, where ethics and IC are in a constant dynamic of transformation and adaptation to generate business results and socio-economic growth of society, because its influence on organizational management incorporates the internal and external environment of the company. Itself and the use of ICT, signifying the progress of society (Bakhsha, Afrazeh and Esfahanipour, 2018).

In particular, ethics, IC and the use of ICTs have currently fostered a significant innovative digital economy for business ethics in its generation processes (training) and in its implementation (organizational culture). This interrelation is visualized in the interaction between economic entities, the production and commercialization processes of companies and the people linked to them. This is how transactional ethics (Corporate Social Responsibility), participatory ethics (democracy in decision-making and implementation of ethics policies in the organizational culture) and recognition ethics (corporate image of the company based on values) emerge (Carrillo, Bensusán, and Micheli, 2016).

Undoubtedly, the increase in IC management capacities and the generation of values depends on the use of knowledge in the organization. If this is positive, business results improve the leadership position and sustainability of the organization, but if it is negative, the organization becomes unproductive isolating itself from its objectives and goals, closing its doors quickly. It is for this reason that companies implement training, training and development policies associated with the mission, vision and values of the organizations, spreading the corporate culture inside and outside the company as a strength of entrepreneurial capacity (Gómez, Londoño and Mora, 2020).

Then, IQ is characterized by the productivity of its human talent and its relationship with knowledge, skills, qualifications, values and experiences. The generation of IC value is reflected in the contributions of training and in the performance and productivity of people, producing positive results (strengths) and reducing negative results (weaknesses) of the organization. In addition to this, ethical values, habits and social skills within the corporate culture and practices produce the knowledge to work

effectively, under a harmonious interrelation that allows quality and good service among social actors (workers, clients, suppliers, shareholders, others) linked to the company (Axtle y Acosta, 2017).

Therefore, IC and ethics are recognized both in people and in the organization, affecting: the corporate image, management systems, production and marketing processes, market positioning and internal and external relationships of the company. In line with the above, knowledge within a company is valued from its IQ and ethics is appreciated in the development of the organizational culture, which is why it is a valuation of tangible assets for their effects on business results and results. prospective returns (Gálvez, Borras, Abadía, 2020).

This connectivity between ethics and IQ in the dynamism of learning organizations has built the complex interconnection of intelligent societies. Indeed, the development of the ability to learn from human talent throughout the organizational structure and the construction of a company based on that knowledge to expand new capabilities is what the notion of intelligent organizations means. The foregoing merges learning, knowledge transfer and the innovative and transformative capacity of people and the organization (Hernández, Muñoz and Jiménez, 2015).

This is how the path towards intelligent societies (IS) is opened, constituted by a community that generates advanced information where organizations develop their ethical values and IC using ICT to essentially prosper in society, that is, the development of People and organizations optimize the quality of life of the population because it increases the value and productivity of their work and their skills in society. It is because of these indications that intelligent societies base learning on people, organizations, governance (harmonious interrelation between the State, citizens and the market) and on the lifestyle with the use of ICT (Jiménez, 2018 y León, 2020).

Therefore, the term intelligent is associated with the adoption and use of automated learning technology, where organizations develop their productivity through ICT and the social actors that interact with them maintain a dynamic of interaction taking advantage of the potentialities and benefits technological. Thus, an IS benefits from the potential of technology to obtain productive citizens, allowing access to resources, applications and networks of interest, increasing their well-being and quality of life in society (Ovalles, Carvajal, Chaustre, Espinoza, Sepúlveda and González, 2018).

All these aspects are related, because societies are made up of people and productive and social organizations that develop and transfer knowledge, transforming the social well-being of people who live and work, that is, this interrelation is a physical and virtual circle of the daily life of people with their social environment. An example of this is the teleworking boom in Europe that has been noticeable for the last 10 years representing 4% of the working population, increasing slowly until this year, expanding sharply to 34% due to confinement due to COVID - 19. Certainly, in Spain teleworking has represented an innovative alternative for companies to meet the demands of the competitive market, enhancing marketing, sales, customer service, advertising, among others.

However, this new working modality has brought with it the need for preventive measures to mitigate health risks, due to the physical and psychological stress to which people are exposed. This critical knot has revealed the existing link between the activities and functions carried out in teleworking and the potential risks associated with it. In other words, the new work environment of preserving the health of the worker and in turn generating the adaptation and optimal functioning of their functions related to communication through the use of high technology communication systems.

In correspondence with the above, there are studies that confirm the existence of health risks in the tasks performed by the teleworker related to: ergonomic conditions, natural biorhythms and work hours, potential distractors, indifference in social relationships, among others. The above has

generated endless effects on the physical and mental health of the worker, evidencing work stress, insomnia, eating disorder, changes in moods, sedentary lifestyle, among others.

As can be seen in the example, the IS transforms the way of life and evolves to the extent that IC is managed efficiently. The promotion of new representations of connectivity and interrelations in the digital environment, together with the possibilities of technological interconnection in people's daily lives, promote IS. This use is generated in technology, digital equipment, social networks and connected devices, meaning the fundamental pillars in communications, applications and services, e-government, connection of systems, among others. (Alarcón, Álvarez, Goyes y Pérez, 2012).

Given these assertions, the formation of a technological culture, learning and knowledge of people in relation to ICT; the potential of government organizations to implement digital leadership; the transformation to technological adaptation of organizations and the incorporation and expansion of digital platforms in competitive markets has made the IS to develop jointly with the IQ and ethical values.

The IS include the use of technology in various areas: political, administrative, economic, social, public services, health care, industrial processes, education, among others. Therefore, KM, organizational culture and people are also part of IS development because they participate and adapt in it. They are involved from the IC perspective because knowledge is promoted for the development of society and all that information is transferred by digital means for the political, economic and social progress of a country. A second example would be the strong growth worldwide, both in the volume of Internet users and in the number of commercial websites and the advertising investment in the network, which is why it is currently considered a mass communication medium (Hernández, Muñoz and Jiménez, 2015).

Precisely, electronic commerce represents a new way of doing business on the Internet without the need to make large investments and to be able to do it directly from the site where the user is (seller and buyer), as long as they have an internet connection. For this reason, collaborative work within an organization and the transactions carried out between online communities allow the increase in sales and the use of communication channels based on the social web and the different platforms, mobile devices, marketing and e-commerce.

In this sense, organizations can simultaneously access their audience and, in turn, adapt their offer to the individual characteristics and needs of their potential customers. The foregoing involves in the process of distribution and online marketing the ability to focus, monitor and measure, which will facilitate a whole learning process for the organization regarding the way its current consumers interact in the digital environment (access to web pages, time spent on it, searches performed, preferences, among others), allowing strategic decisions to be made adapted to the potential market.

From this point of view, the digital environment offers applications and opportunities for the strategic, operational and functional development of marketing, which when combined with electronic commerce reveals a digital scenario of commercial interaction between companies and their potential customers. From the foregoing, it follows, a direct contact between the company, the workers and the client or public interested in the product or service, where the distribution and marketing channel is carried out without intermediaries, allowing to mitigate costs and improving the final sale price. . This implies a CI, that is, an individual (CH), organizational (CE and CR) knowledge of interactive communication that allows communicative bidirectionality between the company and the consumer (Hernández, Muñoz & Jiménez, 2015).

So, for societies to develop intelligently, it is necessary for citizens and organizations (public and private) to integrate their demands and needs into learning and knowledge of technological

innovation, to transform products and services, promote responsibility and social development. , and increase productivity and social welfare. For this reason, societies are currently reforming their technology investment policies towards the use of technology and digital tools for socio-productive purposes, encouraging their adoption and use in organizations and individuals (International Telecommunication Union, 2019).

Seen in this way, the use of technology in societies facilitates sustainable development and its use will depend on the intelligence of organizations and people to enhance technology at the service of progress. It should be noted that management systems and efficiency in the use of resources, the transfer of knowledge and the exchange of information; Organizational strategies and collaborative learning have made it possible to mitigate the digital divide in societies.

In fact, ICTs have influenced the population so much that job opportunities require technological skills in qualification and training programs are currently developed on technological platforms for conducting distance studies. In this sense, the maintenance, transformation, updating or increase of ICT in a country is associated with the characteristics of the development of society, organizations and their citizens (Jiménez, 2018 y León, 2020).

Under all these arguments, knowledge in organizations becomes CI when it understands their social, productive and technological environment. Therefore, it becomes intelligent when knowledge is productive, useful and a generator of value in the organization and in its social environment. For this reason, IC returns to an intelligent organization when it produces, generates and transmits knowledge to be properly applied in society together with technological development, generating social transformations and possibilities for progress and social welfare.

From this viewpoint, IC, ethics and IS are interconnected in a single learning generator of information, interconnection and competences, which involves all parts of society, as a means of development and progress of the citizens of a country. This perception requires an awakening in societies, in citizens and in collaborators and actors inside and outside organizations.

CONCLUSIONS

An intelligent society must be nourished by technological interconnection in all phase of its creation process. In addition, this technological connection is endowed with human talent that generates knowledge (skills, abilities, values...) that can be used to generate differential competitive advantage. This human talent (intellectual capital) is in charge of discriminating the management of ethics, social and business values; all this, counted in turn with the growth of the organization.

KEYWORDS: intellectual capital, knowledge management, endogenous, exogenous, technological connection.

REFERENCES

Alarcón, M; Álvarez, S; Goyes, J y Pérez, O. (2012). Estudio y Análisis del Capital Intelectual como Herramienta de Gestión para la Toma de Decisiones. *Revista del Instituto Internacional de Costos* (10), 49-65. Retrieved from http://www.revistaiic.org/articulos/num10/articulo3_esp.pdf

Axtle, M y Acosta, J. (2017). Measurement And Management Of Intellectual Capital In Higher Education Institutions. *Dimensión Empresarial,* 15(2), 103-115. http://doi.org/10.15665/rde.v15i2.1306

Azofra, A; Ochoa, M; Prieto, B y Santidrián, A (2017). Creando valor mediante la aplicación de modelos de capital intelectual. *Innovar,* 27 (65), 25-38. https://doi.org/10.15446/innovar.v27n65.64887

Bakhsha, A., Afrazeh, A., y Esfahanipour, A. (2018). Identifying the Variables of Intellectual Capital and its Dimensions With the Approach of Structural Equations in the Educational Technology of Iran. Eurasia *Journal of Mathematics, Science and Technology Education*, 14 (5), 1663-16882. https://doi.org/10.29333/ejmste/85037

Carrillo, J., Bensusán, B y Micheli, J. (2016). El Debate Sobre Innovación Y El Progreso Sociolaboral en Covarrubias, S; Sandoval, B; Bensusán, B y Arteaga; J (Eds.), La Industria Automotriz En México: Relaciones De Empleo, Culturas Organizacionales y Factores Psicosociales México: Am Editores.

Chiavenato, I. (2015). Comportamiento Educativo. La Dinámica Del Éxito En Las Organizaciones. Mcgraw-Hill.

Contreras, J., Y Rodríguez, T. (2018). Capital Intelectual Y Ética Gerencial En Las Organizaciones. En Moran, L. (Ed.), Memorias Arbitradas. Jornadas De Investigación Transdiciplinarias (pp. 384-392). Universidad Nacional Experimental Rafael María Baralt.

Docasal, M. (2016). Un Procedimiento Para Medir El Capital Intelectual Y El Desempeño Superior Del Capital Humano En Empresas Hoteleras En Cuba. *Revista Ciencia Y Tecnología*, 11, 32-41.

Farah, A., y Abouzeid, S. (2017) The Impact Of Intellectual Capital On Performance: Evidence From The Public Sector. *Knowledge Management & E - Learning: An International Journal*, 9 (2), 225-238. Retrieved from: http://www.kmel-journal.org/ojs/index.php/online-publication/article/view/373

Gálvez, A; Borras, F & Abadía, J (2020) Indicadores de Gestión del Capital Intelectual para la Banca Comercial Cubana. Revista Retos de la Dirección 2020; 14(1), 310-336. Retrieved from: http://scielo.sld.cu/pdf/rdir/v14n1/2306-9155-rdir-14-01-310.pdf

Gómez, L., Londoño, E., y Mora, B. (2020). Modelos De Capital Intelectual A Nivel Empresarial Y Su Aporte En La Creación De Valor. *Revista Cea*, 6(11), 165-184. https://doi.org/10.22430/24223182.1434

Hernández, H, Muñoz, D, y Jiménez, A (2015). Gestión de la Información Empresarial en las Organizaciones Inteligentes. Universidad Autónoma del Caribe.

International Telecommunication Unión (2019) Un Enfoque Holístico Para Crear Sociedades Inteligentes. Comisiones de Estudios. Retrieved from: https://www.itu.int/dms_pub/itu-d/oth/07/17/D07170000020003PDFS.pdf

Jiménez, L (2018) El Capital Humano e Intelectual como Catalizador de la Gestión Organizacional. *Revista Mundo Fesc,* 15 (1), 83–89. Retrieved from: https://www.fesc.edu.co/Revistas/OJS/index.php/mundofesc/article/view/255/416

León, A (2020) Las Dimensiones del Capital Intelectual y la Cultura Empresarial en las Microempresas del Sector Manufacturero. *Revista Universidad, Ciencia y Tecnología*, 24 (100), 04 – 10. Retrieved from: https://www.uctunexpo.autanabooks.com/index.php/uct/article/view/297/528

Ovalles, L; Carvajal, P., Chaustre, D., Espinoza, S., Sepúlveda, Y y González, J. (2018). Contribución De La Ética Ambiental Y Empresarial A Las Organizaciones. Mundo Fesc, 8(15), 62-72. Retrieved from: https://www.fesc.edu.co/revistas/ojs/index.php/mundofesc/article/view/253

Villegas, E., Hernández, M y Salazar, B (2017). La Medición del Capital Intelectual y su Impacto en el Rendimiento Financiero en Empresas del Sector Industrial en México. Contaduría y Administración, 62(1), 184-206. https://doi.org/10.1016/j.cya.2016.10.002

INTERPRETABILITY CHALLENGES IN MACHINE LEARNING MODELS

Gabriel Marín Díaz, Ramón A. Carrasco González, Daniel Gómez González

Universidad Complutense de Madrid (Spain)

gabriel.marin@ucm.es; ramoncar@ucm.es; dagomez@estad.ucm.es

ABSTRACT

Decisions based on Machine Learning (ML) algorithms are having an increasingly significant social impact; however, most of these systems are based on black box algorithms, models whose rules are not understandable to humans. On the other hand, different public and private organisations, as well as the scientific community, have recognised the problem of interpretability, focusing on the development of interpretable models (white box) or on methods that allow the explanation of black box models.

The aim of this article is to propose a review of the historical evolution and current state of Machine Learning algorithms, analysing the need for interpretability. In this sense, the challenges of interpretability will be addressed from different points of view: in the field of research, legal, industry and regulatory bodies.

INTRODUCTION

Can machines think? This question was posed by Alan M. Turing (1950) in the middle of the 20th century. The answer to this question is the proposal of the so-called Turing test. In this test, Artificial Intelligence (AI) is considered to be the way of acting that imitates the intelligent behaviour of human beings. From then until now, AI has been surpassing human beings in tasks for which intelligence was supposed to be required: strategy games such as chess, driving vehicles, composing symphonies, automatic planning, and a long etcetera that seems to have no end in sight. In fact, the changes that have taken place in recent decades in the telecommunications sector, accompanied by the development of information storage and processing capacity, have led to a paradigm shift that has been given the name of Industry 4.0.

AI corresponds to a field of knowledge that includes Machine Learning (ML) and Deep Learning (DL). In both fields, to solve a problem, models are trained to learn the problem in question from existing data. Once the rules are obtained, we can apply them to new data sets to produce the appropriate answers by applying the rules learned from experience. To perform ML processes, at least three fundamental parts are necessary: input data, the expected results and the measurement of the algorithm's performance so that the algorithm's work can be adjusted through feedback processes (Casella et al., 2013).

An ML model, once implemented, can complete a task much faster and more reliably than any human, delivers consistent results "reliably" and can be infinitely replicated. Training a person to perform a task with the same efficiency is costly and can take years.

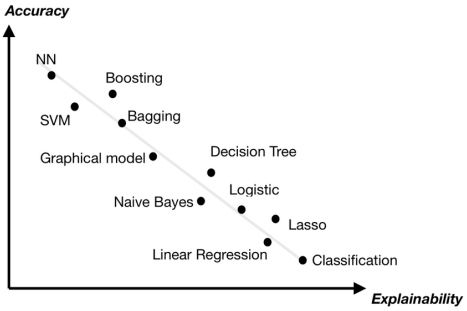
An important aspect of the use of ML is the interpretability of the models once they have been trained. From this point of view some authors distinguish two types of models (Liu et al., 2016):

White-box models are models whose predictive or pattern-identifying behaviour can be clearly explained based on the variables involved.

Black box models are models whose rules are not understandable in a simple way for humans, it would be very difficult to explain how the system came to a certain decision on a certain input. For example, Artificial Neural Networks (ANNs) and DL algorithms in general, such that millions of operations are needed to describe a deep neural network, and there is no way to understand the model in its entirety, hiding how the machine solves a task in increasingly complex models. (Liu et al., 2016).

Some authors even question the interpretability of white-box algorithms (Z. C. Lipton, 2018). Figure 1 shows, as a general rule, that the higher the interpretability of the ML algorithm, the lower its degree of flexibility and consequently the lower its degree of reliability. In other words, there is currently no doubt that the most powerful algorithms are not interpretable.

Figure 1. Interpretability vs Flexibility of ML algorithms.



Source: (Duval, 2019)

In low-risk environments it may not be relevant to understand why a decision has been taken. However, on most occasions, human beings should understand why a decision that affects them individually or collectively has been made. Examples include: a loan decision, a medical decision, self-driving cars, a selection process for a particular job...

In a study conducted in 2019 by Brandon Fornwalt, Geisinger Medical Center, Pennsylvania, they trained two AI algorithms capable of predicting the risk of death in the first year by reading electrocardiograms, even from apparently "normal" people; the algorithm's accuracy was 85% (Samad et al., 2019).

It has been shown that ML models learn very well from the data, but they also pick up biases that may be built into the training data voluntarily or unintentionally. This can make the training model potentially sectarian and discriminate against certain groups and individuals. These potential biases are a key point in investigating the problem of interpretability (Miller, 2019).

In this paper we will review the challenges facing the problem of interpretability in ML models, according to the following structure: in section 2 we will trace the historical evolution of interpretability models, in section 3 we will address the importance of decision-making in this context and how interpretability is a determining factor in the trustworthiness of ML models, and in section 4 we will address the challenges in the field of research, legal, industry and regulatory bodies. In section 5 we will review the interpretability indicators, identifying the quantitative and qualitative factors that make an ML model interpretable, and finally we will draw conclusions.

HISTORICAL DEVELOPMENT OF INTERPRETABILITY

From a historical point of view, in 1950 Turing created the test that bears his name, in 1952 Arthur L. Samuel created the first algorithm capable of learning, in 1956 the concept of Artificial Intelligence was born, in the 70s pattern recognition algorithms emerged, in the 80s expert systems based on rules appeared, the concept of ML began to gain relevance in the 90s being currently one of the most popular subfields within AI, closely linked to mathematical statistics.

Historically, the focus of AI research has shifted towards the implementation of algorithms and models focusing on predictive power to the detriment of interpretability. Model interpretability was emphasised in early machine learning research. The 1970s and 1990s saw the emergence of initiatives such as MYCIN (Britannica, 2018), GUIDON (Clancey, 1987). From the 1980s to the 1990s, systems for tracking alternative lines of reasoning (TMS) were developed. In the 1990s, initiatives emerged in the context of explaining neural networks in healthcare. In 2010, concerns about bias in AI decision-making led to a demand for transparent artificial intelligence and a focus on the interpretability of ML models.

In addition, during recent years we have seen the expansion of social networking systems, which are underpinned by the speed of information processing, communications and storage capacity. As a consequence, due to the exponential increase in heterogeneous data collection and the enormous amount of computational power, machine learning (ML) systems are present in our lives, achieving higher predictive performance and, for most of them, greater complexity (Carvalho et al., 2019).

In practice, what we want is for algorithms to be explainable, i.e. that their operations can be understood by human beings. Despite the correspondence between the two terms, interpretable vs. explainable, there are authors who develop a certain differentiation between the two concepts (Rudin, 2019). Initiatives such as Explainable Artificial Intelligence (XAI) (Gunning et al., 2019), focusing on the interpretability of machine learning algorithms, aims to move towards an interpretable AI model.

On the other hand, and concerning the characteristics that should be attached to interpretable models (Molnar, 2019), we can highlight: the explanations should be contrasting (P. Lipton, 1990), the question we ask ourselves is why a certain prediction was made rather than another, we need to understand by comparison. Furthermore, explanations are selected, i.e., from the set of causes that can give a certain explanation, we are used to selecting one or two causes as the ones most linked to the explanation. Explanations are social, they are part of an interaction between the explainer and the receiver of the explanation where in many cases the social environment is involved. Explanations focus on the abnormal (Kahneman, 1981), causes that are attributed with high potential but low probability. Explanations are true, good explanations prove to be true, the event should be predicted with the highest possible probability. Explanations are consistent with prior beliefs, this is called confirmation bias, devaluing explanations that do not match your beliefs (Nickerson, 1998). Good explanations are general and probable, in the absence of an abnormal scenario, general causes are good explanations (Gaussian curve).

As can be seen, the concept of interpretability suggests the involvement of more than one area of knowledge (Carvalho et al., 2019), at least three stand out: data science developing predictive models, social sciences leading to understanding, and human-machine interaction to empower the user (Abdul et al., 2018).

Interpretability is therefore a necessary milestone for the success of ML and AI itself. As stated by (Roy, 2017), "In the end, mathematical models should be our tools, not our masters", which is only possible with interpretability.

Interpretation methods for machine learning can be classified according to several criteria (Z. C. Lipton, 2018).

Intrinsic or post hoc? This criterion distinguishes whether interpretability is achieved by restricting the complexity of the machine learning model (intrinsic) or by applying methods that analyse the model after training (post hoc). In the former case, interpretability is inherent to the model, in the latter case, the methods may or may not be decoupled from the ML model.

Especific or agnostic? Interpretation tools are limited to specific model classes e.g. linear regression, or applicable to any model once trained (post hoc). Agnostic tools can be used on any machine learning model and separate the explanation from the type of model. They offer the freedom to choose a set of models to address a problem and then compare them.

Local or global? Does the interpretation method explain an individual prediction or the entire behaviour of the model? Or is the scope somewhere in between?

IMPORTANCE OF INTERPRETABILITY IN ML

According to Miller "Interpretability is the degree to which a human can understand the cause of a decision" (Miller, 2019). This means that in the interpretation of a model there is a directly proportional link to causality, understanding why a prediction was made by the model.

On the other hand, a correct prediction only partially solves the original problem, an explanatory black box model that has 85% agreement with the original model does indeed explain the original model most of the time, however, it is wrong 15% of the time. Therefore, confidence in this black box model is limited to that 85% reliability. The higher the interpretability of a machine learning model, the easier it is for someone to understand why certain decisions or predictions have been made. In some cases it may not be relevant to understand why a certain decision has been made, especially in a low risk environment (Doshi-Velez & Kim, 2017). In most cases, however, we need to understand why, as it can help us to better understand the problem and the reasons why a model may fail.

The trend in recent years has been to take advantage of the characteristics of ML and in particular the predictive power of black box algorithms for high-risk decision making, for example in legal, financial and health services, all of which have a profound impact on society and in particular on human lives (Rudin, 2019), a fundamental point in the investigation of the problem of its interpretability (Miller, 2019; Molnar, 2019).

The following are some examples of biases applied by ML algorithms in different technologies and areas:

On 7 November 2019, Ruby on Rails creator and entrepreneur David Heinemeier Hansson shared a disturbing story on Twitter (Business, 2019), alleging that the Apple card was discriminating against his wife. Both he and his wife applied for this card, but he received a credit limit 20 times higher than she did, even though they applied for the card at the same time, and filed joint tax returns.

In the autumn of 2019, Google unveiled a Machine Learning technology called BERT (Bert, 2019) to improve its search criteria by incorporating the context of words accompanying the search object. However, the data it works with corresponds to the largest digital library in history, bringing with it decades of biases and prejudices that are built into the search algorithm, and are likely to be perpetuated.

In 2015, the Google Photos app labelled two African-Americans as "gorillas". (BBC Mundo Tecnología, 2015). Google engineers analysed the account and discovered that the algorithm had problems adjusting for photo contrast, lighting and skin tone. In addition, they confessed that, due to this same problem, the algorithm labelled white-skinned people as dogs and seals.

In 2016, some of LinkedIn's algorithms were found to have a gender bias (Day, 2016), recommending better paid jobs for men. This casuistry may be reinforced by the fact that high-paying jobs are predominantly held by men.

In 2016 Microsoft launched "Tay" (BBC Mundo, 2016), a chatbot whose purpose was to mimic the behaviour of a curious teenager seeking to engage in casual conversation on social media with a target audience of 18-24 year olds. In less than 24 hours, Tay, through tweets, showed her empathy for Hitler or her support for genocide by answering questions from social media users.

In 2016, the COMPAS algorithm (Correctional Offender Management Profiling for Alternative Sanctions) to predict recidivism, developed by Northpointe (now Equivant), was accused of bias against African Americans (Larson et al., 2016).

In 2018, oncologists criticised IBM's Watson for Oncology for providing unsafe and inaccurate recommendations (Ross, 2018).

In 2018, Amazon's resume screening system was found to be biased against women (Dastin, 2018).

In 2019, algorithms behind Apple's credit card are accused of gender bias (Fast Company, 2019).

As can be seen, there are a significant number of predictive models whose outcome determines a negative impact on people's safety and rights, leading to serious violations of ethical and equity principles. In this context, it is essential to build tools that allow for model exploration, in particular to explain the model, examine and evaluate its performance, and understand its weaknesses and failures.

Equally important are algorithmic audits to detect discrimination and bias, and to incorporate ethical values into these systems (Carvalho et al., 2019).

According to (Doshi-Velez & Kim, 2017) the aspects that could be optimised through interpretability are as follows:

Impartiality, unbiased, non-discriminatory predictions.

Privacy, protection of information.

Reliability, small changes in the data input do not affect the prediction.

Causality, only collect causal relationships (cause-effect).

Confindence, systems must explain their decisions in order to be reliable.

The fundamental objective is to gain trust and social acceptance of ML algorithms through interpretability.

INITIATIVES TOWARDS INTERPRETABLE AI

Currently, there is no real consensus on what interpretability in Machine Learning models is, nor is it clear how to measure interpretability. However, technology companies, international organisations and public administrations are aware of the problem and are taking steps to mitigate the consequences of discriminatory bias in algorithms.

Technology Companies

IBM launched the Fairness 360 Kit project in 2018 (Hughes et al., 2020), this open source toolkit helps to examine, report and mitigate discrimination and bias in machine learning models. Adversarial Robustness 360 (ART) Toolbox (Adversarial Robustness Toolbox, 2021) is a Python library for machine learning security. AIX360 Toolkit (AI Explanability 360, 2021) is a comprehensive open source toolkit with various algorithms, codes, guides, tutorials and demos that support the interpretability of machine learning models.

Microsoft has a model interpretation SDK in Azure Machine Learning for use in Python (Microsoft, 2021).

Google has an API, Explainable AI (Hughes et al., 2020), is a set of tools and frameworks capable of helping to debug and understand the behaviour of Machine Learning models.

H2O Driverless AI, a machine learning platform (H2O.ai., 2020) offered by H2O.ai, offers interpretability as one of its distinguishing features.

DataRobot (DataRobot, 2021), is another commercialised ML solution, "includes several components that result in models that are highly interpretable by humans".

Google Vizier is a service for optimising black box models. (Golovin et al., 2017).

Facebook, in collaboration with Georgia Tech, published an article showing a tool for visual exploration of industrial-scale DNN models (Kahng et al., 2018).

Uber recently announced Manifold, a model-agnostic visual debugging tool for ML (Zhang et al., 2018).

Other companies are taking steps in the same direction; however, perfection cannot be expected, there will always be undetected biases, or biases that cannot be eliminated.

Legislation, Organisations and Regulatory Documents

As business and government decisions become increasingly automated, the need to protect against black box algorithms will be critical. We will need to know how and why decisions are made, understanding is crucial to move forward safely. To this end, it will be necessary to work on the control and auditing of algorithms whose decisions directly affect people, independent bodies will be needed that are capable of determining the "quality" of the algorithm, providing sufficient guarantees to citizens, thus increasing the social acceptance of this type of practice. Ensuring that the following qualities are met: fairness, privacy, reliability, robustness, causality, trustworthiness (Doshi-Velez & Kim, 2017).

Profiling and automated decisions can pose significant risks to individual rights and freedoms. European and Spanish data protection legislation obliges and requires certain safeguards. Article 22 of the GDPR (UE, 2016) provides that European citizens have the right not to be subject to a decision based solely on automated means, including profiling, if the decision produces legal effects which significantly affect them in a similar way.

On the other hand, we have the ISO/IEC 27001 standard (Blackmer, 2018), which aims to ensure the confidentiality, integrity and availability of an organisation's information and the systems and applications that process it, this standard has been developed by the International Organisation for Standardisation (ISO), and will have to adapt its content to the needs arising from the interpretability in the IA (Weller, 2019).

One of the most notable entities in the field of AI research, Defense Advanced Research Projects (DARPA), created the XAI programme (Gunning et al., 2019). In 2016 the White House Office of Science and Technology Policy (OSTP) published the US report on AI entitled "Preparing for the Future of Artificial Intelligence". (Bundy, 2017).

The Royal Society, which is the UK Academy of Sciences, published a report on its machine learning project in April 2017 (Royal Society of Great Britain, 2017).

In Spain, the technical subcommittee for standardisation CTN 71/SC 42 - Artificial Intelligence and Big Data was set up in December 2019 (UNE, 2021) precisely to elaborate standards in the field of AI, participating in the development of global standards being developed in the international committee ISO/IEC JTC 1/SC 42 Artificial Intelligence (Standardization, 2021).

In April 2018, the European Commission published the following communication on Artificial Intelligence for Europe (Commission, 2018). In 2019, the High-Level Expert Group on Artificial Intelligence formulated guidelines on trustworthy AI (European Commission, 2019). In parallel, the first coordinated plan on AI was published in December 2018 as a joint commitment with the Member States (Digitales et al., 2020).

The Commission's White Paper on AI, published in 2020, sets out a clear vision for AI in Europe: "an ecosystem of excellence and trust that lays the foundation for today's proposition" (Comisión Europea, 2020).

In April 2021, the European Commission, in coordination with member states and with the aim of strengthening trust and excellence in AI, launched a risk-based approach that penalises, and even bans, AI systems that are considered a clear threat to security. High-risk systems will be subject to strict obligations before they can be placed on the market (Munchen, 2021).

Gradually, both EU and non-EU countries will join such initiatives, so as to offer a glimmer of hope and try to ensure reliability in AI systems.

Science

The easiest way to achieve interpretability is to use interpretable ML algorithms (white box models), including linear regression, logistic regression, decision trees, RuleFit and Naive Bayes. (Molnar, 2019). From these models, features can be extracted in terms of other features that allow the model to be defined and interpreted at a global level (Sundararajan et al., 2017).

On the other hand, there are model-specific methods of explanation, many of which are designed to be used with neural networks that are difficult to interpret (black box models). Another option is to extract knowledge from a more complex model by approximating it with an interpretable model (Bastani et al., 2017; Tan et al., 2018).

Finally, we have the agnostic methods of explanation, which do not depend on the ML model, and are post hoc, the great advantage of these models over the specific ones is their flexibility.

An overview of agnostic models is represented in the table 1 (Carvalho et al., 2019):

The current trend is to focus on model-independent interpretation tools; it, is much easier to automate interpretability if we separate the interpretation method from the model used. With agnostic methods we can replace both the learning model and the interpretation method, the capabilities provided by this system are highly scalable. (Carvalho et al., 2019; Ribeiro et al., 2016; Molnar, 2019).

Explanation Method	Scope	Result
Partial Dependence Plot	Global	Feature Summary
Individual Condition Expectation	Global / Local	Feature Summary
Accumulated Local Effects Plot	Global	Feature Summary
Feature Interaction	Global	Feature Summary
Feature Importance	Global / Local	Feature Summary
Local Surrogate Model	Local	Surrogate Interpretable Model
Shapley Values	Local	Feature Summary
BreakDown	Local	Feature Summary
Anchors	Local	Feature Summary
Counterfactual Explanations	Local	(new) Data Point
Prototypes and Criticisms	Global	(existent) Data Point
Influence Functions	Global / Local	(existent) Data Point

As we have seen in this document, interpretability is not only a scientific question, other areas of knowledge linked to the human being are involved. The need for interpretability is inherent to the desire to know, to the human being's need to answer the question of why.

INTERPRETABILITY INDICATORS

Can we measure and evaluate interpretability? Despite all the work being done in different areas of knowledge, this question unfortunately remains unanswered. However, the work that is being done is oriented along two clear lines: the use of ML algorithms that allow a high degree of precision and making the decision adopted by these systems interpretable, explainable to human beings.

A review of the literature suggests that little work has been done to develop models to measure and evaluate interpretations, so that the most appropriate explanation can be chosen (Honegger, 2018). However, we can distinguish between two types of indicators when comparing and evaluating explanations (Carvalho et al., 2019), quantitative and qualitative.

Among the qualitative indicators (Doshi-Velez & Kim, 2018) suggests the following questions:

What are explanations composed of? Which features are predominant in an explanation?

How many subsets of blocks of features can an explanation be made up of, and if we remove any blocks, is the result affected?

How are these blocks formed? What composition should be given between the blocks?

What relationships might be more intuitive to humans?

Are any random processes part of the explanation?

For quantitative indicators (Sundararajan et al., 2017), (Honegger, 2018) established a framework for measuring the consistency of explanatory methods whose prediction must be consistent with human explanation. It is necessary according to (Honegger, 2018) to relate the object (instance and prediction) to its subsequent explanation (importance value of features).

Identity. Identical objects must have identical explanations. If a method of explanation is asked to explain a certain object, the explanations it gives must be the same.

Separability. Non-identical objects cannot have identical explanations. It follows from the previous premise.

Stability. Similar objects must have similar explanations. If slight perturbations considerably modify the response, the system is not stable.

In addition, other variables must be considered, such as completeness, the audience has to verify the validity of the explanation. Correctness, the explanation must generate confidence. And finally, compactness, the explanation must be precise, brief, and concise.

CONCLUSIONS

And at company and individual level, what can be done? We propose to intervene on the following aspects.

Addressing biases

While, as we have said, the task will not be easy, we do not know how black box algorithms work, but we can act on the biases so that the decisions taken by the algorithms are aligned with the rights of the people.

Digital maturity

Machine learning is the subject of many expectations, but are companies ready for data governance? Science is constantly developing machine learning tools, but can they be integrated into a company's business processes? Most companies have grown based on technological silos, integrating pieces with little or no scalability. There is no such thing as a single piece of data; the fundamental task of data scientists is to "find out" where the information is to be able to analyse and make predictive models. Their core business has more to do with an Agatha Christie novel than with analysis and predictive modelling useful for the business. The expectations generated by the media and the occasional guru about AI and its application are unlikely to be fulfilled until the business culture of the short-term changes.

Machine learning will grow, not at the speed it is touted, but slowly and steadily. Fundamental to this is the process of business digitalisation that starts from a single data model, from the integration of all the company's information to be able to extract working models where AI can develop. Better formulas are needed to integrate AI into the business processes of companies, perhaps it would be useful to develop machine learning tools that are easy to use and can be automatically integrated with business management processes, this would help to make a technological leap in the digitisation process and will be the first step from childhood to youth.

The next step towards maturity could be the adoption of a full automation model of business management processes, tasks could be posed as decision problems solved by machine learning.

Interpretability as a catalyst

At this point, interpretability will be critical to ensure that algorithms are responsive to reality, trying to minimise the impact of biases. On the other hand, transparency is the norm in any organisation - decisions need to be supported by an understanding of the underlying tasks. The interpretability of algorithms must be fundamental to trust in predictive black box models. If we use interpretation-agnostic methods, we can automatically apply them to any model that emerges in a machine learning process and train surrogate models that improve the predictions. If we are able to do this, we may be able to improve our understanding of intelligence and become better at creating intelligent machines.

The opportunities are obvious, but so are the associated dangers. In an increasingly anumerical society (Hand & Paulos, 1992) where decision making is usually done through System 1 thinking: quick, intuitive and emotional (Kahneman, 2012), versus System 2 thinking: slow, deliberative and logical. We can sense that this speed, immediacy of the everyday and short-termism can invade us, leaving the decisions that require thought and meditation to a third party (the machine).

There is room for improvement and progress, the expectations are good, but so are the challenges!!!!

KEYWORDS: Machine Learning, Interpretability, Deep Learning, Bias, Artificial Intelligence.

REFERENCES

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). 17. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*. https://doi.org/10.1145/3173574.3174156

Adversarial Robustness Toolbox. (2021). *Adversarial Robustness Toolbox*. https://adversarial-robustness-toolbox.org/

AI Explanability 360. (2021). *AI Explanability 360*. https://aix360.mybluemix.net/

Bastani, O., Kim, C., & Bastani, H. (2017). 137. Interpreting blackbox models via model extraction. *ArXiv*.

BBC Mundo. (2016). *Tay, la robot racista y xenófoba de Microsoft*. Bbc. https://www.bbc.com/mundo/noticias/2016/03/160325_tecnologia_microsoft_tay_bot_adolescente_inteligencia_artificial_racista_xenofoba_lb%0Ahttp://www.bbc.com/mundo/noticias/2016/03/160325_tecnologia_microsoft_tay_bot_adolescente_inteligencia_artificial_raci

BBC Mundo Tecnología. (2015). *Google pide perdón por confundir a una pareja negra con gorilas*. Bbc. https://www.bbc.com/mundo/noticias/2015/07/150702_tecnologia_google_perdon_confundir_afroamericanos_gorilas_lv

Bert, G. (2018). *Google BERT*. https://cloud.google.com/tpu/docs/tutorials/bert

Blackmer, W. S. (2018). 84. EU general data protection regulation. *American Fuel and Petrochemical Manufacturers, AFPM - Labor Relations/Human Resources Conference 2018*, *2014*(April), 45–62. https://doi.org/10.1308/rcsfdj.2018.54

Britannica, E. (2018). *MYCIN*. https://www.britannica.com/technology/MYCIN

Bundy, A. (2017). 20. Preparing for the future of Artificial Intelligence. *Ai & Society*, *32*(2), 285–287. https://doi.org/10.1007/s00146-016-0685-0

Business, C. (2019). *Apple co-founder Steve Wozniak says Apple Card discriminated against his wife*. https://edition.cnn.com/2019/11/10/business/goldman-sachs-apple-card-discrimination/index.html

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). CAT. A - Machine learning interpretability: A survey on methods and metrics. *Electronics (Switzerland)*, *8*(8), 1–34. https://doi.org/10.3390/electronics8080832

Casella, G., Fienberg, S., & Olkin, I. (2013). An Introduction to Statistical Learning. In *Springer Texts in Statistics*. http://books.google.com/books?id=9tv0taI8l6YC

Clancey, W. J. (1987). The GUIDON Program. *MIT Press Series in Artificial Intelligence*.

Comisión Europea. (2020). Libro Blanco sobre la Inteligencia Artificial - un enfoque europeo orientado a la excelencia y la confianza. *Comisión Europea*, 1–31. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf

Commission, E. (2018). *Artificial Intelligence for Europe - Communication*. https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF

Dastin, J. (2005). *Amazon scraps secret AI recruiting tool that showed bias against women*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

DataRobot. (2021). *DataRobot*. https://www.datarobot.com/wiki/interpretability/

Day, M. (2016). *How LinkedIn's search engine may reflect a gender bias*. The Seattle Times. https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias/

Digitales, S., Unidos, E., Europa, H., & Digital, P. E. (2020). *Los Estados miembros y la Comisión colaborarán para impulsar la inteligencia artificial « fabricada en Europa » Contexto Más información*. 2019–2021.

Doshi-Velez, F., & Kim, B. (2017). *41. Towards A Rigorous Science of Interpretable Machine Learning*. *Ml*, 1–13. http://arxiv.org/abs/1702.08608

Doshi-Velez, F., & Kim, B. (2018). *152. Considerations for Evaluation and Generalization in Interpretable Machine Learning*. 3–17. https://doi.org/10.1007/978-3-319-98131-4_1

Duval, A. (2019). *Explainable Artificial Intelligence (XAI) Explainable Artificial*. *April*. https://doi.org/10.13140/RG.2.2.24722.09929

European Commission. (2019). *COM(2019) 168 final Building Trust in Human Centric Artificial Intelligence*. 11. https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

Fast Company. (2019). *I applied for an Apple Card. What they offered was a sexist insult*. https://www.fastcompany.com/90429224/i-applied-for-an-apple-card-what-they-offered-was-a-sexist-insult

Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., & Sculley, D. (2017). 8. Google vizier: A service for black-box optimization. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Part F1296*, 1487–1496. https://doi.org/10.1145/3097983.3098043

Goodman, B., & Flaxman, S. (2017). 88. European union regulations on algorithmic decision making and a "right to explanation." *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). 18. XAI-Explainable artificial intelligence. *Science Robotics*, *4*(37), 0–1. https://doi.org/10.1126/scirobotics.aay7120

H2O.ai. (2020). *H2O Driverless AI*. https://www.h2o.ai/products/h2o-driverless-ai/

Hand, D., & Paulos, J. A. (1992). Innumeracy: Mathematical Illiteracy and its Consequences. In *Applied Statistics* (Vol. 41, Issue 1). https://doi.org/10.2307/2347643

Honegger, M. R. (2018). *79. Shedding Light on Black Box Machine Learning Algorithms*. *August*.

Hughes, R., Edmond, C., Wells, L., Glencross, M., Zhu, L., & Bednarz, T. (2020). *eXplainable AI (XAI)*. 1–62. https://doi.org/10.1145/3415263.3419166

Kahneman, D. (1981). *The Simulation Heuristic*.

Kahneman, D. (2012). Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011. In *Etc* (Issue October).

Kahng, M., Andrews, P. Y., Kalro, A., & Chau, D. H. P. (2018). 39. ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 88–97. https://doi.org/10.1109/TVCG.2017.2744718

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How We Analyzed the COMPAS Recidivism Algorithm*. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lipton, P. (1990). Contrastive explanation. *Contrastivism in Philosophy*, 11–34. https://doi.org/10.4324/9780203117477

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 35–43. https://doi.org/10.1145/3233231

Liu, H., Cocea, M., & Gegov, A. (2016). Interpretability of computational models for sentiment analysis. *Studies in Computational Intelligence*, *639*(March), 199–220. https://doi.org/10.1007/978-3-319-30319-2_9

Microsoft. (2021). *Instalar el SDK de Azure Machine Learning para Python*. https://docs.microsoft.com/es-es/python/api/overview/azure/ml/install?preserve-view=true&view=azure-ml-py

Miller, T. (2019). 95. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Book*, 247. https://christophm.github.io/interpretable-ml-book

Munchen, T. U. (2021). European approach to Artificial Intelligence. *E-Conversion - Proposal for a Cluster of Excellence*, 29–50. https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Zeitschrift Für Neurologie*, *199*(1–2), 145–150. https://doi.org/10.1007/BF00316552

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Model-Agnostic Interpretability of Machine Learning*. *Whi*. http://arxiv.org/abs/1606.05386

Ross, C. (2018). Watson for Oncology. *STAT*, 1–30. papers3://publication/uuid/5566F158-417A-46D3-B583-04EE273812A1

Roy, M. (2017). 80. Cathy O'Neil. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishers, 2016. 272p. Hardcover, $26 (ISBN 978-0553418811). In *College & Research Libraries* (Vol. 78, Issue 3). https://doi.org/10.5860/crl.78.3.403

Royal Society of Great Britain. (2017). 24. Machine learning : the power and promise of computers that learn by example. In *Report by the Royal Society* (Vol. 66, Issue January).

Rudin, C. (2019). 9. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Samad, M. D., Ulloa, A., Wehner, G. J., Jing, L., Hartzel, D., Good, C. W., Williams, B. A., Haggerty, C. M., & Fornwalt, B. K. (2019). Predicting Survival From Large Echocardiography and Electronic Health Record Datasets: Optimization With Machine Learning. *JACC: Cardiovascular Imaging*, *12*(4), 681–689. https://doi.org/10.1016/j.jcmg.2018.04.026

Standardization, I. O. (2021). *ISO*. International Organization for Standardization. https://www.iso.org/committee/6794475.html

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *ArXiv*.

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). 77. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 303–310. https://doi.org/10.1145/3278721.3278725

UE. (2016). *Artículo 22 UE RGDP*. https://www.privacy-regulation.eu/es/22.htm

UNE. (2021). *UNE Normalización Española*. https://www.une.org/encuentra-tu-norma/comites-tecnicos-de-normalizacion/comite/?c=CTN 71/SC 42

Weller, A. (2019). 85. Transparency: Motivations and Challenges. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11700 LNCS*(Section 2), 23–40. https://doi.org/10.1007/978-3-030-28954-6_2

Zhang, J., Wang, Y., Molino, P., Li, L., & Ebert, D. S. (2018). 40. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *ArXiv*, *25*(1), 364–373.

FACT CHECKING AGENCIES AND PROCESSES TO FIGHT AGAINST FAKE NEWS

Eglée Ortega-Fernández, Graciela Padilla-Castillo, Sonia Laura Carcelén-García, Mario Arias-Oliva

Antonio de Nebrija University and Complutense University of Madrid (Spain),
Complutense University of Madrid (Spain) Madrid (Spain); Complutense University of Madrid (Spain)¸
Complutense University of Madrid (Spain)

eortegafe@nebrija.es; gracielp@ucm.es; sonialca@ucm.es; mario.arias@ucm.es

INTRODUCTION

We are living in the Smart Society, where citizens are connected to other people and devices that generate an incalculable amount of information each second. The Smart Society looks for the increase of the well-being of citizens, the strength of the economy, and the effectiveness of institutions based on the innovative use of emerging smart digital technologies (Chakravorti & Chaturvedi). According to the World Economic Forum (2020), the digital universe is expected to reach 44 zettabytes in 2020.

Citizens are producing, consuming and exchanging information every day: sharing photos, personal videoconferences, social media content, or streaming services. And business is generating, exchanging and consuming digital information every day with eCommerce, teleworking, intranet access or global information systems. But paradoxically, at the moment with the most information available for citizens, we are probably worse informed than ever before. There are several reasons that explain this paradox.

We are losing the freedom to choose our source of information. An algorithm is deciding for us, filtering and suggesting what is the information of any kind (such as commercial, news, media or streaming) that we should consume. That way of organizing information provokes a narrowed mind in which vision and understanding of individuals and society are biased. The shared vision and the shared values are the ones decided by the algorism with the ethical concerns that this fact represents: who manages the algorithm, manages society and individuals.

It could be considered as the hacking of the human brain. According to Harari (2019), to hack a human being three things are needed: solid knowledge of biology, a lot of data about human behaviours and an important computer capacity. The Inquisition or the KGB never succeeded in penetrating human beings because they lacked this knowledge of biology, data about individual and aggregate behaviours, and computer capacity.

Nowadays companies such as the GAFAM (Google, Amazon, Facebook, Apple and Microsoft) and governments are developing all these tools and knowledge, and probably they will be -or they are- hacking us. They will not only be able to predict our decisions, but also to manipulate our feelings. In this digital environment, governments and corporations are using this new technique to manipulate us, and fake news is an emerging strategy.

The term "fake news" is used to mean disinformation through the use of dissemination of totally false messages or manipulation of information (misinformation) trying to bias information to achieve the desired goal (Watts, 2018). In both cases, the intention of cheating to serve a certain cause exists. Satirical information or news messages misleading that based on a true fact biases it by manipulation or decontextualization are increasingly in our society. Fake news is spreading in politics; journalism, law and policies or social media (Zimdars & McLeod, 2020).

These messages can have geopolitical consequences, for example by creating or fuelling conflicts, changing the intention to vote, or by putting public opinion for or against a certain social problem. Citizens increasingly trust these types of messages, being not able to distinguish between what is fake news and real news.

According to Arias-Oliva and Khawly (2021), the control of messages disseminated through the Internet is increasing their social and economic influence. The news that is created and shared quickly in social media can generate significant advertising returns when readers access the original website (Fielden, Grupac & Adamko, 2018), and fake news with polarized messages generate more traffic than real news.

A study found that fake news are diffused significantly faster, deeper, and more broadly than the real news, and the effects were more pronounced for political fake news than for other information such as terrorism, natural disasters, science, urban legends, or financial information (Vosoughi, Roy & Aral, 2018). Fake news are designed to become viral information, exploring all possible aspects to attract the reader's attention, from the title design to the language used throughout the body of the text (Baptista & Gradim, 2020).

Beyond fake news practices as an unethical method to increase traffic and revenue, an even more concerning practice is the use for geopolitical purposes. Facebook and Twitter are no longer used mainly as a way to connect us with family and friends or connecting brands with customers. Facebook now has 2.7 billion accounts worldwide and has become a political influencer. Twitter, with 1.3 billion accounts created, has only 330 million active accounts per month, with a significant number of manipulated accounts (EC Financial News, 2020).

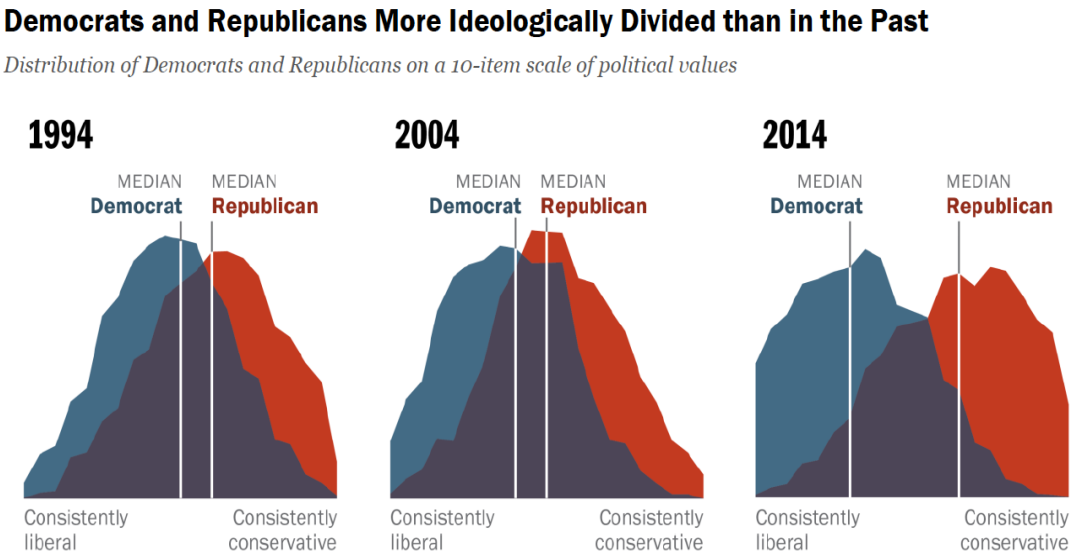
Both networks have become a powerful instrument of institutional and political communication, but also of disinformation. Through real or fictitious users, they are instruments to manipulate public opinion. Armies of manipulated Twitter accounts are used through complex artificial intelligence systems to redirect public opinion for certain purposes. These systems are called bots, which from the so-called "bots farms" try to influence the electoral results representing a serious threat to the sovereignty of countries.

Marcellino, Johnson, Posard and Helmus (2020) analysed these techniques in the last presidential elections in the United States. Their report proves the existence of electoral interference through the use of trolls (fake people who spread a variety of sensitive information and news in an exaggerated way) and super-voters (highly networked accounts that can spread messages effectively and quickly).

The study did not identify the origin of the accounts, but interference seems to favour Russian interests according to the authors, who transparently describe the methods used to identify the dubious accounts.

As an example of this algorithm polarization is the report published by Pew Research Center (2014). We can see the consequences of and fake news in the USA political arena. According to the study, 92% of Republicans are to the right of the median Democrat, and 94% of Democrats are to the left of the median Republican, as we can see in Figure 1.

Figure 1. Polarization of USA political opinions.



Source: Pew Research Center (2014).

FACT CHECKING NEWS AGENCIES AND PROCESSES

Fact checking news agencies are groups of specialists, who are usually journalists, dedicated to denying or reviewing fake news to show the truth to the public. This type of communication media and projects dedicated to fact checking have grown in number around the world, boosted by the infodemic that the COVID-19 pandemic has created (Ortega-Fernández, 2021).

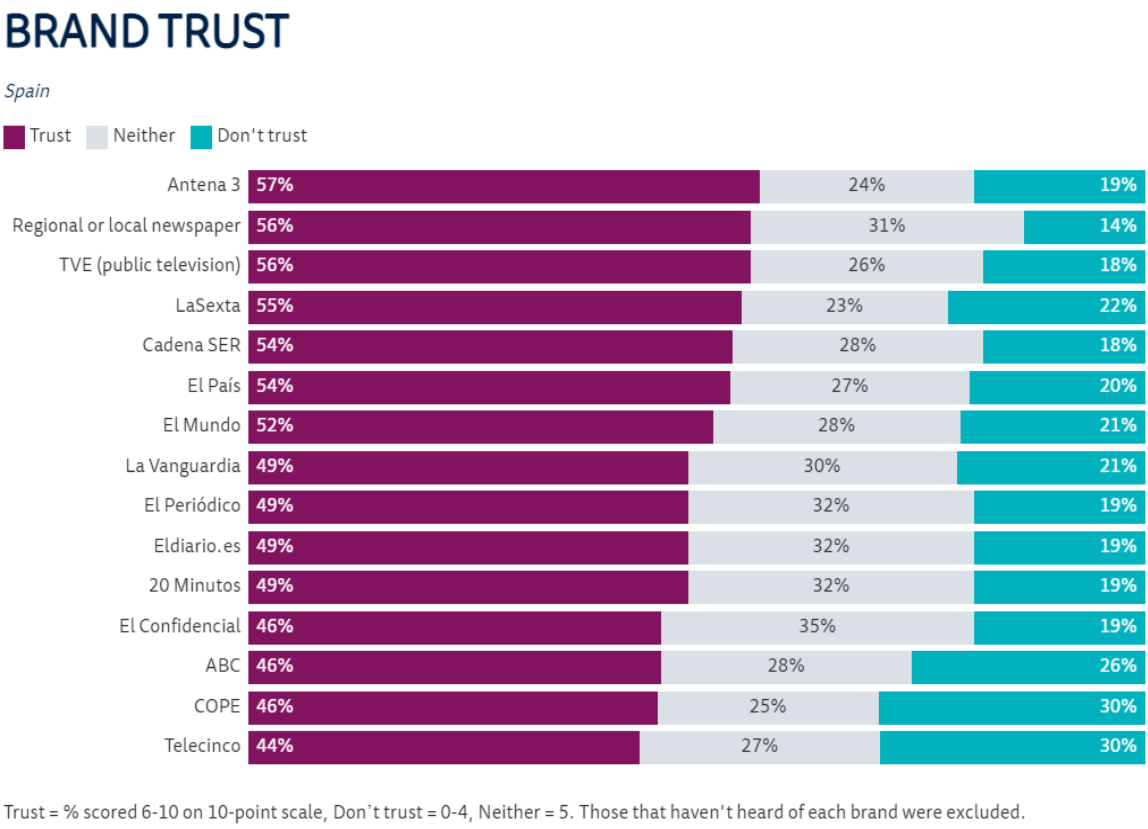
In Spain these groups have emerged and grown in an incipient way since 2018, when the digital verifier Newtral was born, followed by Maldita.es and EFE Verifica. After these three pioneers, other media have emerged in our country that are not only dedicated to verification, but also incorporate in their newscasts or web pages a section or section dedicated to disproving the hoaxes that run through social networks or media.

However, when studies are carried out on trust in these media dedicated to fact checking, they are not reflected like the conventional media. Perhaps due to its incipient creation and dissemination, people continue to resort to traditional media, even when reliability levels are just above 50%.

Fact checking news agencies focus on very diverse criteria and tools to be able to determine the factors that make up a fake news. A study endorsed by the Andrés Bello Catholic University of Venezuela and the Venezuelan Fake News Observatory (Pabón and Viloria, 2020) has recently been published where tools aimed at verifying images, text, geolocation or cameras are listed and briefly described, web, applications, browser extensions, online platforms and online courses, for this purpose.

As they are digital tools, some new ones are being updated and created even at the time of this publication, so we will focus on highlighting the procedure that is carried out by the media / platforms that participated in this research.

Figure 2. Confidence levels in Spanish media.



Source: Digital News Report (2020)

The processes are assimilated and adjusted according to the requirements and processes established by the agencies. In Venezuela there is a digital media completely dedicated to verification, called EsPaja.com (Figure 3), a verification portal that was born in October 2019 as an initiative of Transparency Venezuela with the sponsorship of the European Union.

Ana Griffin, editor-in-chief, details that the process can begin in two ways: editorial decision or by user requests. User requests come through WhatsApp or directly through the page. In both cases, the editor-in-chief takes note of what was requested and selects them according to the audience's interest, giving preference to issues that have to do with Venezuela or with public services. When they define it by editorial means, they are guided by the follow-up or continuity of the information in the news field.

Once the topics are defined, the drafting process begins to generate a note where it is defined if the information received is true, false or if it is between truth and lies, using verification tools for images, reverse search in Google, tools such as Invid or TinEye, even when required, a geolocation analysis is incorporated through Google Maps.

Figure 3. News agency homepage: Espaja.com.

Source: www.espaja.com (2021)

When we review this work to verification in other countries, the process changes. The fact checking news agency called Factual (Figure 4), that belongs to Agence France-Presse (AFP), have an active team since 2017, who verify news in Latin America with 20 journalists who work between languages: Catalan, Portuguese and Spanish.

The coordinator of the team, Elodie Martínez, details that the process is more proactive, monitoring through social networks, in addition to assessing the content of the requests they receive through WhatsApp. They rely on the CrowdTangle tool to weigh the virality and urgency of the issues, giving priority to issues that could cause damage to health, society, attacks or conflicts. They use their own extension co-developed by the agency, Invid WeVerify, which allows to launch the reverse search in 6 engines at the same time.

Another important point is that they try to contact the author of the original publication, in order to try to reach the official source, or consult various sources of their own to contrast the information.

Figure 4. News agency homepage: Factual.afp.com.

Source: www.factual.afp.com/list (2021)

The fact checking portal Verifikado.com (Figure 5), according to its director Fernando Nunez-Noda, it is more automated. It begins when their own system that they have developed collects information from the news ecosystem based on consolidated generalist media and news integrators such as Google News to have a first filter, which reduces work time.

In the second step, the system determines the news of doubtful origin and reviews the media that replicates the information to analyze the internal positioning factors of those who publish and other external factors that include the public transparency of the medium that publishes, the quality that has and the recognition of who writes the news, to mention a few factors.

Regarding the content, they perform an analysis of factors called forensic to establish whether it is news that could be presented as evidence in a United States court of law, based on the concept of forensic evidence for the judicial system of that country. These steps allow them to categorize the news while continuing to feed the system to establish the patterns that constantly guide searches.

It emphasizes that the false news that is generated in China has different characteristics from those that are produced in Cuba, to mention an example, so the system they use usually analyzes these patterns and traces backwards, such as the IP (Internet Protocol) of the news, which can determine the origin of the news and its entire journey.

In Verifikado they are developing an algorithmic system based on Big Data, for the detection, disassembly and monitoring of fake news through an application.

Figure 5. News agency homepage: Verifikado.com.

Source: www.verifikado.com (2021)

From Argentina, there are also proposals focused on the verification of news with specific methods. The Desconfío.org (Figure 6) project aims to promote research on the dynamics of disinformation in the Spanish language and train journalists and media with strategies to detect and stop disinformation, offering digital solutions to users.

The Distrust Method, as they call it themselves, offers 10 steps for the verification of suspicious content, among which is included the reverse search of images, the identification and consultation of the cited sources, in addition to the contrast of the alternative sources, to mention some of the parameters they recommend.

This decalogue is recommended by the platform in the training it provides and is taken as a starting point to verify any type of news that has overtones of falsehood or is of doubtful origin, finally classifying it as Wrong Content (Mis-Information), Misinformation (Dis-Information) or Harmful Information (Bad-Information).

Figure 6. News agency homepage: Desconfio.org.

Source: www.desconfio.org/ (2021)

When reviewing these references, added to the avalanche of information that citizens receive every day, we might think that it is not possible for ordinary people to comply with these very specific processes. On the other hand, the fact checking media themselves declare that they cannot cover everything.

If we focus on the role of social networks as vehicles and amplifiers of fake news, it is perhaps the most difficult to control. This type of news appeals to the emotions of the users and awakens not very rational and very impulsive reactions because people seek to reinforce their beliefs, convictions and justify their fears.

Taking into account that the verifications contradict news that are popular, even if they are not real, people are frustrated by having fallen into the trap of sharing them and that is why they do not usually spread the denials with the same vehemence that they distribute false information that is amplified, and they go viral very easily on social networks.

However, platforms such as Twitter and Facebook have developed functionalities that seek to limit the spread of doubtful information. There is also a bot for WhatsApp, developed by the International Fact-Checking Network (IFCN) that allows the user to verify the information. YouTube has also joined in this drive to generate reliable information, as well as Google's section, called Discover, which relies on artificial intelligence.

Recognizing that there are key themes that are repeated in fake news, the global panorama that includes the pandemic, polarization in politics, struggles for equality and the environment, among others, is a very powerful breeding ground for keep creating and spreading fake news.

Fake news is usually part of campaigns organized by countries, economic groups, politicians or even free agents affiliated with some ideology or specific movements, which have a defined intention, either for political or economic interests.

Trends indicate that agencies will move from content curation to algorithm curation, since artificial intelligence may be the key to addressing the huge amount of data that arises from the virality of fake news.

CONCLUSIONS

Fake news has existed for a long time, what is relevant is the current moment, is that issues of great importance (pandemic, struggles for equality, political polarization) come together with citizens and hyperconnected people who are exposed to newspapers to hundreds of thousands of information from all.

Without a doubt, we need to reflect on the processes that must be followed and the options that exist to verify the information that circulates, since many factors conspire against a healthy news system.

Technology facilitates this task with tools that are being updated, but in a double direction, this same technology, through social networks, also makes it easier for hoaxes to go viral and reach millions of people in a very short time. We live in the post-truth era, where people share what is good for their cause or that supports their ideology, sometimes without criteria and without stopping to think if it is real.

This same technology is what also allows the development of systems through Big Data and the conjunction man / machine, whether artificial intelligence helps us with the task of eradicating fake news.

The current situation presents us with a challenging panorama, but as can be seen, detailing the actions that are being carried out, there are initiatives that are constantly working to end this phenomenon. It is a professional profile that opens opportunities for researchers and journalists to develop in the field of fact checking.

The creation of detection patterns, developed by individuals and by social media, plays a fundamental role in these actions, since they could be used in a generalized way with the same speed with which the information is received.

KEYWORDS: fake news, media reliability, social media, disinformation, misinformation.

REFERENCES

Arias-Oliva, M. & Khawly, N. (2021). *La redefinición de las geo-estrategias internacionales* in Padilla, G., *DEFENCERCA: Acercar la Defensa a la ciudadanía y a los comunicadores. Retos y posibilidades*. Fragua.

Baptista, J. P., & Gradim, A. (2020). Understanding Fake News Consumption: A Review. *Social Sciences*, 9(10), 185.

Bhaskar C., Chaturvedi R.S. (2017). The “Smart Society” of the Future Doesn’t Look Like Science Fiction. *Harvard Business Review*, 10. Retrieved from https://hbr.org/2017/10/the-smart-society-of-the-future-doesnt-look-like-science-fiction

Charlie Beckett (2017) Truth, trust and technology. *Media Asia, 44*(2), 98-101. https://doi.org/10.1080/01296612.2017.1455571

CE Noticias Financieras (2020). Social media and democracy. CE Noticias Financieras. Retrieved from https://login.bucm.idm.oclc.org/login?url=https://www.proquest.com/wire-feeds/social-media-democracy/docview/2464600767/se-2?accountid=14514

David S. Morris, Jonathan S. Morris & Peter L. Francia (2020) A fake news inocu-lation? Fact checkers, partisan identification, and the power of misinformation. *Politics, Groups, and Identities*. https://doi.org/10.1080/21565503.2020.1803935

Edelman. (2018). Trust Barometer. Global Report. https://tinyurl.com/yambnm9x

Edson C. Tandoc Jr., Zheng Wei Lim & Richard Ling (2018) Defining "Fake News". *Digital Journalism, 6*(2), 137-153. https://doi.org/10.1080/21670811.2017.1360143

Fielden, A., Grupac, M., & Adamko, P. (2018). How users validate the information they encounter on digital content platforms: the production and proliferation of fake social media news, the likelihood of consumer exposure, and online deceptions. Geopolitics, History and International Relations, 10(2), 51-57. http://doi.org/10.22381/GHIR10220186

Harati Y. N. (2019). Los cerebros 'hackeados' votan. *El País* (Jan. 6, 2019). Retrieved online from https://elpais.com/internacional/2019/01/04/actualidad/1546602935_606381.html

Marcellino, Johnson, Posard & Helmus (2020). Foreign Interference in the 2020 Election: Tools for Detecting Online Election Interference. Santa Monica, CA: RAND Corporation, 2020. Retrieved from https://www.rand.org/pubs/research_reports/RRA704-2.html

Ortega-Fernández, E. (2021). Verificación de la información. Procesos y Fact Checking en tiempos de infodemia y COVID-19. *Comunicando en el siglo XXI: Nuevas fórmulas. Colección Comunica*. Editorial Tirant lo Blanch. ISBN: 978-84-1853436-2.

Pew Research Center (2014). Political Polarization in the American Public How Increasing Ideological Uniformity and Partisan Antipathy Affect Politics, Compromise and Everyday Life. Pew Research Center. https://www.pewresearch.org/politics/wp-content/uploads/sites/4/2014/06/6-12-2014-Political-Polarization-Release.pdf

Reuters Institute (2020) Digital News Report 2020. https://tinyurl.com/y7sqj3ve

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151.

World Economic Forum (2020). How much data is generated each day? World Economic Forum. Retrieved from https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

Zimdars M. & McLeod K. (eds) (2020). *Fake News: Understanding Media and Misinformation in the Digital Age*, The MIT Press: Cambridge, MA.

SOCIAL ROBOTS: MAIN ETHICAL CHALLENGES AND ISSUES

Áurea Subero-Navarro, María Alesanco-Llorente, Jorge Pelegrín-Borondo y Mario Arias-Oliva

Universidad de La Rioja (Spain), Universidad de La Rioja (Spain), Universidad de La Rioja (Spain), Universidad Complutense de Madrid (Spain)

aurea.subero@unirioja.es; maria.alesanco@alum.unirioja.es; jorge.pelegrin@unirioja.es; mario.arias@ucm.es

INTRODUCTION

In recent years, the great advances that have been made in different disciplines such as computer science and mechanics have enabled the development of robots created not only for the realization of industrial tasks, but also to interact with people in healthcare environments, providing different services such as care of the elderly people, advisory tasks in commercial environments, medical tasks, etc. (Torras, 2014). These types of robots are called social robots and their rise has boosted the interest of researchers.

Today the implementation of robots is taking place practically in every area of society. However, the discussion over the use of robotics and artificial intelligence (AI) has increased as the possible consequences of the use of robots for the economy, employment and society are beginning to be seen (Huang and Rust, 2018). These future changes in society generate both, expectations and fears (Mick and Fournier, 1998). That is why in a society where AI is becoming more relevant, it is necessary to define a number of limits in relation to the mode of use of AI and the way of interaction with human beings (Santos González, 2017).

Technoscientific and technological development implies the introduction of improvements in human life but also implies the emergence of numerous risks in which it is necessary to influence. One of the relevant issues to address is the impact of ethics on technoscience, that is to say, how to apply ethics to emerging new technologies. This is because any cultural change requires, in turn, an ethical rethink to avoid unwanted situations in the future (Valls Prieto, 2019).

It is therefore necessary not only to anticipate the possible scenarios that may occur but also to identify the future moral problems that will arise from them. There are numerous initiatives and projects related to roboetics that aim to answer all these questions (Torras, 2014).

The future will be conditioned by the decision-making that takes place at this time and not only by the legislator, but also by each of the agents who have the capacity to influence them. It is therefore necessary to regulate both individual and collective behaviour in relation to minimum requirements and to carry out a periodic evaluation of it. How the sharing of benefits and costs is appreciated will obviously influence the blessing of consumers. The introduction of these technological developments does not necessarily imply inequality, as long as these technological tools are subject to values and standards (Grau Ruiz, 2019).

Another main challenge of ethics or theft is based on the use of language. Currently, the terms used can be applied exclusively to humans and not to robots. Decision-making is completely different and contingencies arise in terms such as "consciousness" or "empathy" that constitute human realities and are mis-applied to machines. It will be necessary to introduce terms that adapt to robotics. The risk of this is very great because if robots are equated to humans, there is a risk that machines will resemble people (Noeo Tech, 2017).

Therefore, it will be essential to implement an ethical system that promotes the ethical behavior of the different agents participating in it. However, the main problem in relation to this issue is that human beings usually know what is ethical, but not how to achieve it, so this supposes the main challenge in the field of Philosophy and Law (Grau Ruiz, 2019).

In this regard, the European Parliament in 2017 approved the robotics report establishing an ethical code of conduct. Its purpose is to serve as: "an ethical guidance framework for the design, production and use of robots". All this has led to the emergence of a new term coined by Gianmarco Veruggio: robotics (Noeo Tech, 2017).

In addition, it is also necessary to influence the opinion of the population in relation to the use of these robots. Ethics reflects the plurality of legal, moral and religious norms that govern in a community (Berger et al., 2008). Reidenchach and Robin (1990) consider that individuals use more than one basis to make ethical judgments so that the employment of multiple dimensions is necessary to capture the meaning of ethical judgment. Therefore, after the revision of literature dedicated to moral philosophy, they developed the Multidimensional Scale of Ethics (MES) which includes five dimensions based on contemporary normative moral philosophies: moral equity, utilitarianism, relativism, selfishness and contractualism.

This theoretical framework is taken as a reference. The objective of the present research is to find out if the ethical action of consumers in relation to the use of robots takes some place in the minds of the consumers involved in it and analyze the main ethical issues of the introduction of social robots in our lives. Based on this theoretical background, the authors propose to advance in the knowledge of the impact of ethical judgment on the intention of using social robots in commerce. To do this, the respondents were asked their opinion about the advantages and disadvantages of using social robots in commerce.

MAIN ETHICAL ISSUES

The robotics industry is growing at a dizzying pace. Social robots are increasingly present in all areas: education, health, care of elderly people, customer service (Beer et al.2011; Conti et al.2019; Alaiad & Thou, 2014) and their adoption rates have accelerated exponentially reaching sales of service robots at annual rates above 30% (International Federation of Robotics, 2018).Mckinsey'report (2018) says that of the nineteen major industries, the retail industry has the greatest potential to create value through the use of autonomous technologies and AI which translates to more than 600 billion dollars annually.

In the light of these facts, it seems important to analyze the ethical issues raised by the use of social robots. These ethical issues regarding the use of robots and their impacts on society are the main subject of roboethics (Demir, 2017). Advanced robotics can create problems if people don't understand the consequences that can result from introducing an increasingly intelligent technology (Alsegier, 2016). Therefore, addressing the key principles of roboethics as they arise is essential to guarantee a correct symbiosis in human-robot interaction (Tsafestas, 2018). The need for ethical considerations in the development of these intelligent systems is becoming one of the main areas of research giving rise to different initiatives such as the IEE on the ethics of autonomous systems, the foundation for responsible robotics and the association on AI, among others (Dignum, 2018).

In this section, the main ethical problems of human use of social robotics will be described and the problems arising from the use of this emerging technology will be examined. The design of these systems is not only relevant in terms of their responsible use (Houkes & Vermaas, 2010), but also requires a responsible design. Among the main ethical dilemmas are:

- **Privacy:** Currently there is a discussion about the privacy and surveillance of information technology (Macnish, 2017), that is, access to private and sensitive data. So, security and data protection has become a relevant issue. The introduction of technology has accelerated exponentially while regulation has taken a long time to respond (Proposal for a Regulation laying down harmonised rules on artificial intelligence-Artificial Intelligence Act, 2021) (European Commission, 2021).

  However, currently robotic systems have not yet played an important role in the area of privacy. Nevertheless, this will change when they are introduced in different scenarios. Data protection is at risk with the rapid development of AI since it's use involves the processing of large amounts of data.

  Therefore, the Spanish Data Protection Agency (AEPD) has published a guide to adapt to the general data protection regulation (RGPD) for products and services that include AI components (AEPD, 2021).

- **Behavior manipulation:** Ethical issues in relation to AI are not only reduced to the use of data, but also include the use of that information to modify and manipulate human behavior (Dezfouli et al.2020). In this sense, a study conducted by Data61 researchers from the Commonwealth Scientific and Industrial Research Organization or CSIRO has concluded that AI can find vulnerabilities within human decision-making (Dezfouli et al.2020).

  According to Steinert, 2014 robots are mere extensions of human capabilities and can be used as tools to modify a situation according to human wishes. Robots are now considered amoral systems since technology is neutral in relation to use (Westerlund, 2020). That is, a robot can be used to perform surgery and save a life or, on the contrary, it can be used to kill someone as a result of human will (Steinert, 2014).

  In fact, robots could be used as killer weapons (Demir, 2017). However, even if a robot is built as an autonomous and intelligent system, the ethical concerns that arise from its usage are linked to the human design (Westerlund, 2020). Some researchers point out that robots are analogous to domestic animals in terms of liability (Kelley et al., 2020). In other words, if a robot is involved in an accident or a problem, the responsibility lies with the owner (Westerlund, 2020). For this reason, the protection of humanity is necessary in relation to protection against possible manipulation and in terms of responsibility.

- **Opacity of IA systems:** opacity and biases constitute the main ethical challenges of AI (Floridi & Taddeo, 2016; Mittelstadt & Floridi, 2016). One of the characteristics of algorithms is the opacity ("black box") (Monasterio Astobiza, 2017). Although they are hidden, they are invisible, because they are inscrutable between the layers and sub-layers of computer programming. They are opaque in the sense that they are hermetic to interpretation (Monasterio Astobiza, 2017).

  So, there is concern regarding the use of AI systems for automated decision support and predictive analytics in relation to the lack of process and community involvement and auditing (Whittaker et al. 2018).

  Because of this, people fail to understand the basis for making a decision. To alleviate this darkness, a new discipline has been created: explainable AI that enables human beings to understand the decisions made (Barredo Arrieta et al.,2019).

For its part, the EU Regulation 2016/679 has taken this opacity into account and provides that when consumers are faced with a decision based on algorithmic processing, they have the right to a legal explanation of the decision.

- **Bias in decision systems:** Automated systems that carry out decisions and predictive analysis operate on the data and, based on it, generate a decision as an output. Algorithms influence important decisions in people's daily lives, so mistakes should be avoided and they should be infallible and ethical. Choi et al., 2010 analyze biases and group them into three broad categories:

  A) Biases derived from problems with the drafting of the question.

  B) Biases derived from problems with the design and layout of the questionnaire.

  C) Biases derived from problems with the use of the questionnaire.

  However, when this problem is limited to the context of AI, it's necessary to specify it.

  According to Hao (2019) bias can occur at different times in the process, although he highlights the three stages where it usually arises: in the definition of the model, in the data collection and in the preparation of data for its use. Therefore, if the algorithm that's going to make the decision of a development is biased, the result will also be biased (Salazar & Escribano Otero, 2021). There are numerous examples of theses biases that have been seen in the press such as the following: "The Amazon algorithm that does not like women". This article refers to the Amazon algorithm that was used in the staff selection that rewarded men over women (Salazar & Escribano Otero, 2021).

  There are constant efforts to remove biases from AI systems. However, they are in the early stages of learning (Brownsword et al., 2017). Technological solutions have their limits in the sense that a mathematical notion of justice is required, which is very difficult to achieve (Selbst et al.2019).

To avoid these ethical dilemmas, the European Parliament has approved a report of guidelines for the use of AI in both, civil and military fields. The European Parliament has urged the European Commission to create a framework of principles and legal obligations on artificial and robotic intelligence. The proposal is based on the fact that AI must be governed by a series of requirements that guarantee security, transparency and accountability with safeguards against bias and discrimination. As well as the right to redress, social and environmental responsibility, privacy and data protection (Madiega, 2019).

METHOD

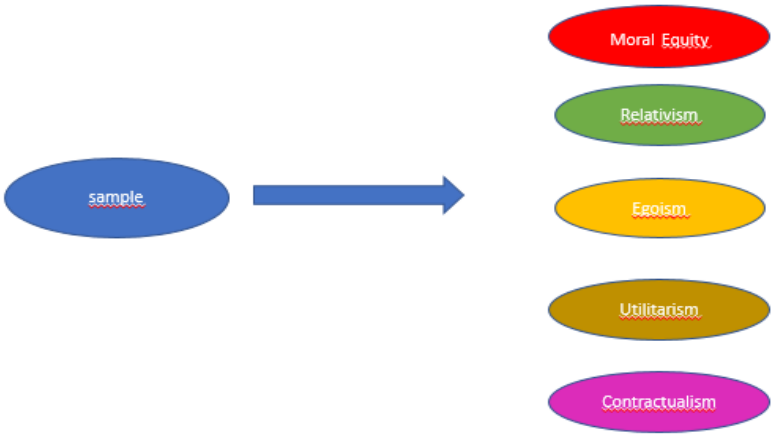
A personal survey has been applied to a sample of 100 individuals over 18 years of age residing in Spain. Regarding the characteristics of the sample, they are: Men: 48% and Women: 52% from 18 years to more than 65 years. The survey was based on two open questions about the advantages and disadvantages about the use of social robots in the commerce.

In terms of mechanics, the survey takers located the participants according to gender and age quotas. The interviews were personal and recorded to ensure the quality of the field work and the information.

Once the readings corresponding to the review of the survey applied to the sample had been carried out, the responses were categorized. In this sense, they were classified into categories according to

the thematic criteria related to the objectives of the research. The categories were the following: moral equity, relativism, selfishness, utilitarism and contractualism (Shawner & Senneti, 2009).

Model 1.



Source: Self-Elaboration.

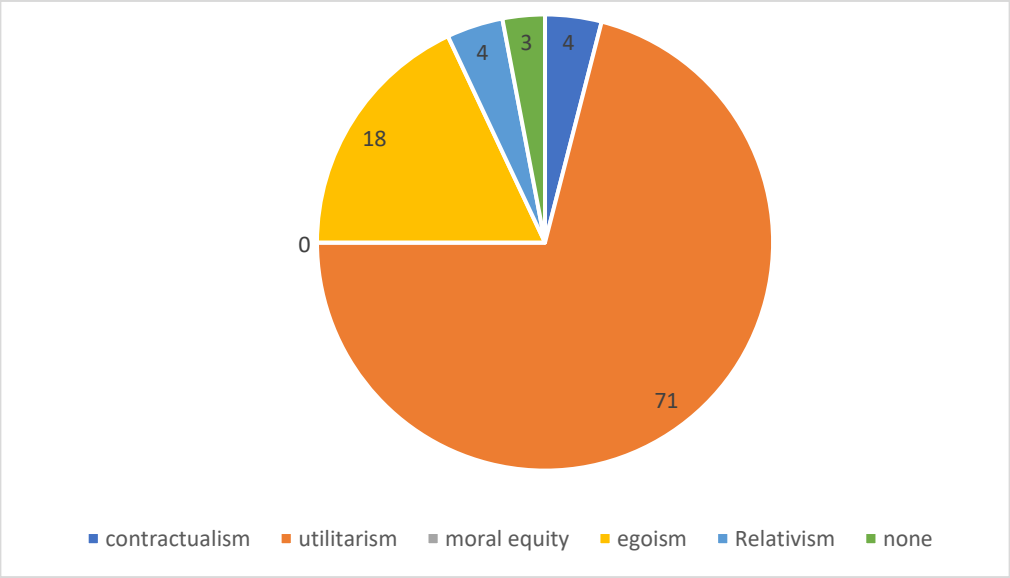
RESULTS

To achieve this goal, one hundred people have been interviewed. These people have been asked open questions about what they consider to be the main benefits and disadvantages of the use of social robots in retail (Shawver and Senneti, 2009). The results show the following:

- Utilitarianism: a dimension based on consequential theories that consider morality to be measured by the consequences of the actions performed (Reidenbach and Robin, 1990). The results show that when claiming advantages of the use of social robots in retail, the respondents made a balance of their costs and benefits and 71% of them were guided by utilitarian reasons such as speed of service, handling large amounts of information and the ability to easily locate products. On the other hand, only 31% alleged utilitarian reasons when alleged inconveniences.
- Moral Equity: dimension based on the theory of justice (Rawls, 1971). The results show that at the time of stating the advantages of the use of social robots in retail, none alleges reasons related to moral equity, while, for their part, in stating the disadvantages 59% alleged these reasons, such as the dehumanization of work, destruction of employment and increase in the number of people in unemployment.
- Selfishness: dimension also based on consequential theories, but focusing exclusively on the individual consequences (Reidenbach and Robin, 1990). The results show that 18% used selfish arguments such as "not having to be kind to dependents". For its part, when it comes to stating inconveniences, none alleged selfish reasons.
- Relativism: it is defined as the "perception of what is correct versus incorrect based on guidelines/parameters of the social/cultural system (Nguyen and Biderman, 2008). The results of the research show that when declaring the benefits of the use of robots in retail 4% of respondents alleged relativistic motivations by comparing the right and wrong of the use. They argued, for example, that it would be good because of the increased of efficiency while it wouldn't be correct because of the destruction of employment. For their part, when they exposed the drawbacks of the use of robots in retail, 9% made relativistic arguments.

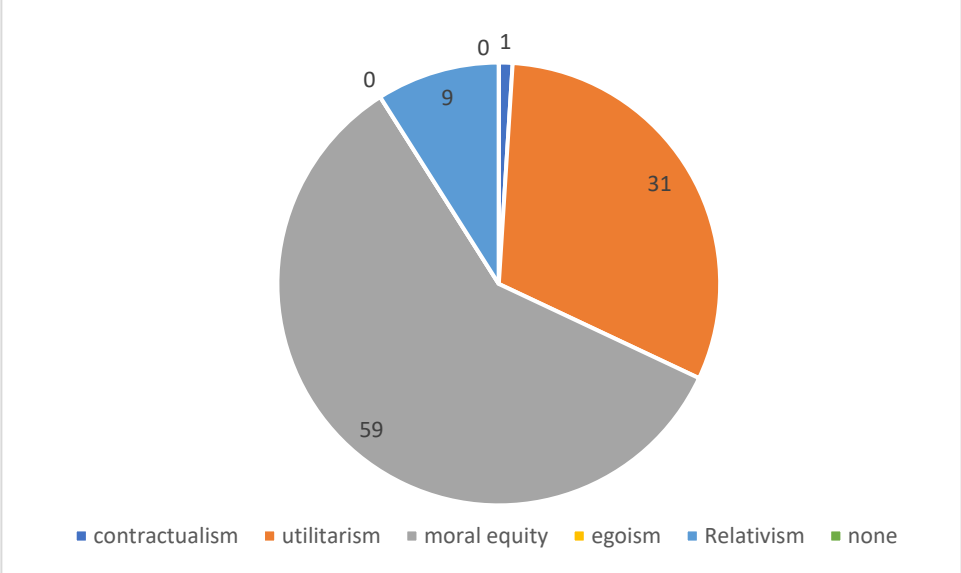
- Contractualism: a deontological dimension that encompasses different notions such as implicit obligation, contracts, duties and rules (Reindenbach and Robin, 1990). The results show that when declaring the benefits of the use of robots in retail 4% claimed contractual reasons, while when it came to stating the inconveniences 1% alleged such reasons.
- For their part, 3% of respondents said they found no benefit from the use of robots in retail.

Graphic 1. Results Advantages.



Source: Self-elaboration.

Graphic 2. Results Disadvantages.



Source: Self elaboration.

The influence of ethical judgment on the intended conduct is a conclusion from the investigation. When respondents were asked about the benefits of using robots, most of them alleged utilitarian reasons. It is clear that one of the strengths of robots are the great possibilities they offer us: they

enable economic growth and the development of traditional businesses, they accelerate and optimize business processes, they are a source of information and recommendation and constitute a source of entertainment for users. However, in exposing the inconveniences of their employment, 59% of people alleged moral equity reasons. This is mainly due to the fear generated by the introduction of robots into our lives due to process automation and job destruction. However, robots can be used with the aim of complementing humans and freeing them from the most routine tasks, rather than replacing humans.

ACKNOWLEDEGEMENTS

This research was funded by the bridge grants for research projects awarded by the University of La Rioja (PROYECTOS PUENTE PP-2020-02), subsidized by Banco Santander, the COBEMADE research group at the University of La Rioja and ayudas para estudios cientificos de tematica riojana del IER año 2020-2021.

KEYWORDS: social robots, ethics, artificial intelligence, ethical judgment.

REFERENCES

AEPD (2021). Requisitos en auditorías de tratamientos que incluyen Inteligencia Artificial. *AEPD*. Retrieved from: https://www.aepd.es/es/documento/requisitos-auditorias-tratamientos-incluyan-ia.pdf.

Alaiad, A., and Zhou, L. (2014). The determinants of home healthcare robots adoption: An empirical investigation. *International Journal of Medical Informatics*, 83(11), 825-840.

Barredo Arrieta, A., Díaz Rodríguez, N., Del Ser, J., Bennelot, A., Tabik, S., Barbado González, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, V.R., Chatila, R., & Herrera, F. (2019). Explanaible Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward responsable AI, Information Fusion, 58, 2-73.

Beer, J.M., Prakash, A., Mitzner, T.L., & Rogers, W.A. (2011). Understanding robot acceptance, technical Report HFA-TR-1103, Atlanta GA: Georgia Institute of Technology School of Psicology. Rerieved from: https://smartech.gatech.edu/bitstream/handle/1853/39672/HFA-TR-1103-RobotAcceptance.pdf;accessed

Berger, C., González-Franco, M., Ofek, E. & Hinckley, K. (2018). The uncanny valley of haptics. *Science Robotics,* 3 (17), 1-2.

Brownsword, R., Scotford, E., & Yeung, K. (2017). The Oxford Handbook of Law, Regulation and Technology, Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199680832.001.0001

Choi, B., Granero, R., & Pak, A. (2010). COMITÉ EDITORIAL Comunicación Especial. Revista Costarr Salud Pública, 19(2), 106–118.

Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P and Malhotra, S. (2018). Notes from the AI Frontier: Applications and Value of Deep Learning. *Mckinsey & Company*. Retrieved from: https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning.

Conti, D., Cattani, A., Di Nuovo, S., and Di Nuovo, A. (2019). Are future psychologists willing to accept and use a humanoid robot in their practice? Italian and English Students Perspective. *Frontiers in psychology*, 10 (2138), 1-13.

Demir, K. A. (2017). Research questions in roboethics. *Mugla Journal of Science and Technology*, 3(2), 160- 165. http://doi.org/10.22531/muglajsci.359648

Dezfouli, A., Nock, R., & Dayan, P. (2020). Adversarial vulnerabilities of human decision-making. Data61, Commonwealth Scientific and Industrial Research Organization (CSIRO), 1-8. Retrieved from: https://www.pnas.org/content/pnas/117/46/29221.full.pdf

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology*, 20 (1-3), 1-3.

European Comission (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence-Artificial Intelligence Act. Retrieved from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELLAR%3Ae0649735-a372-11eb-9585-01aa75ed71a1

Floridi, L. & Taddeo, M. (2016). What Is Data Ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083). http://doi.org/10.1098/rsta.2016.0360

Grau Ruiz, M. A. (2019). La importancia de la ética en el mundo de la inteligencia artificial y la robótica. *Universidad Complutense de Madrid*, 76-77.

Hao, K. (2019). This is how AI bias really happens—and why it's so hard to fix. *MIT Technology Review.* Retrieved from: https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-biasreally-happensand-why-its-so-hard-to-fix/.

Houkes, W. & Vermaas, P.E. (2010). Technical Functions: On the Use and Design of Artefacts, (Philosophy of Engineering and Technology 1), Dordrecht: Springer Netherlands.

Huang, M-H. & Rust, R.T. (2018). Artificial intelligence in service. *Journal of Service Research*, 21(2), 155-172.

International Federation of Robotics. (2018). "World Robotics report 2018", Press conference summary, Tokyo.

Kelley, R., Schaerer, E., Gomez, M., & Nicolescu, M. (2010). Liability in Robotics: An International Perspective on Robots as Animals. *Advanced Robotics*, 24(13), 1861-1871. http:// doi.org/10.1163/016918610X527194.

Macnish, K. (2017). *The Ethics of Surveillance: An Introduction,* London: Routledge.

Madiega, T. (2019). EU guidelines on ethics in artificial intelligence: Context and implementation. European Parliamentary Research Service, 1-13. Retrieved from: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf.

Mick, D.G. & Fournier, S. (1998). Paradoxes of Technology: Consumer cognizance, emotions, ad coping strategies. *Journal of Consumer Research,* 25(2), 123-143.

Mittelstadt, B. D. & Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. *Science and Engineering Ethics*, 22(2): 303–341. http://doi.org/10.1007/s11948-015-9652-2

Monasterio Astobiza, A. (2017). Ética algorítmica: Implicaciones éticas de una sociedad cada vez más gobernada por algoritmos. ILEMATA,24,185-217.

Nguyen, N. T., & Biderman, M. D. (2008). Studying ethical judgments and behavioral intentions using structural equations: Evidence from the multidimensional ethics scale. *Journal of Business Ethics*, 83(4), 627–640.

Noeo Tech (S.f.). Roboética, una necesidad urgente. *Edita Noeo Tech*. Retrieved from: https://noeotech.com/

Rawls, J. (1971/1999). *A theory of justice (revised edition).* Cambridge: Harvard University Press.

Reidenbach, R. E., & Robin, D. P. (1990). Toward the development of a multidimensional scale for improving evaluations of business ethics. *Journal of Business Ethics*, 9(8), 639–653. http://doi.org/10.1007/BF00383391

Santos González, M. J. (2017). Regulación Legal de la Robótica y la Inteligencia Artificial: Retos de futuro.

Salazar, I., & Escribano Otero, J. J. (2021). ¿Necesitan tener ética los robots? UEM STEAM essentials, 1-7.

Selbst, A.D., Boyd, D., Friedler S.A., Venkatasubramanian S., & Vertesi, J. (2019). Equidad y abstracción en sistemas sociotécnicos, en Actas de la Conferencia sobre equidad, responsabilidad y transparencia — FAT*'19, Atlanta, GA: ACM Press, 59–68. http://doi.org/10.1145/3287560.3287598

Shawver, T. J., & Sennetti, J. T. (2009). Measuring ethical sensitivity and evaluation. *Journal of Business Ethics*, 88(4), 663–678. Retrieved from: https://doi.org/10.1007/s10551-008-9973-z

Steinert, S. (2014). The Five Robots - A Taxonomy for Roboethics. *International Journal of Social Robotics*, 6, 249-260. http://doi.org/10.1007/s12369-013-0221-z

Torras, C. (2014). Robots sociales: Un punto de encuentro entre ciencia y ficción. *Revista Mètode*, 1-5. Retrieved from: https://dialnet.unirioja.es/servlet/articulo?codigo=4765325

Tsafestas, S. G. 2018. Roboethics: Fundamental Concepts and Future Prospects. *Information*, 9(6), 148. http://doi.org/10.3390/info9060148

Valls Prieto, J. (2019). El reto de una robótica e inteligencia artificial honesta con las personas. *Infobae*.

Westerlund, M. (2020). An ethical framework for smart robots. *Technology Innovation Management Review*, 10 (1), 35-44. http://doi.org/10.22215/timreview/1312

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Manthur, V., Myers West, S., Richardson, R., Schultz, J., & Schwartz, O. (2018). AI Now Report 2018, AI Now Institute, 1-62. Retrieved from: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

NEUROETHICAL PSYCHOLOGY IN TRANSHUMANISM

Jesús Jiménez Olarte

Universidad de La Rioja (Spain) / Francisca Breton Foundation (Spain)

jesusjimenez@cop.es

INTRODUCTION

Technological progress in the search for solutions to internal conflicts runs parallel to the transition process in terms of the change of the humanist paradigm. Pharmacological solutions allow us to modify our state of mind, nanotechnology begins to be applied to the brain and thus neuroethics is transformed (Bostrom, 2005). The problems of cryopreservation contain within themselves a future solution as the problem is defined, and embryo selection is a real possibility and currently in the process of implementation (Simon, 2006).

The aim of this study is not to question the ethical and moral posthumanist legitimacy of the limits that should or should not be exceeded, but to foresee the psycho-emotional conflicts derived from this symbiotic evolution that will lead to the postulation of moral premises that will propose solutions to the ethical conflicts that will be generated in the future. A future tending towards technological progress applicable to all areas of the human being, both directly related to survival (quality of life and life expectancy) and to adaptation to the environment and to consciousness as individual beings.

We are currently at a transition point towards a transhumanist development in which the possibility and moral tendency to improve the physical, intellectual and psychic capacities of the human species through the application of new technologies and eugenics (Overall, 2017), with the aim of eliminating all undesirable aspects of the human condition such as illness, suffering, ageing, and even death, lead us irremediably to the transhuman species by evolution (Hughes, 2004). Such potentialities open the window to a posthuman being, in which, in parallel, an evolution in moral and social ethical conflicts should be envisaged (Fukuyama, 2004).

Keywords: transhumanism, neuroethics, transhumanist psychology, ethical technology, evolutionary neurosis, posthumanism.

FROM TRANSHUMANISM TO NEUROETHICAL POSTHUMANISM.

Contrary to what many humanist theorists refer to, transhumanism is not a current of thought, or at least it is not ONLY a current of thought derived from parts of science that are joined with paste of ideology. It is the naming of a real, present and cumulative consequence of scientific and technological advances that, for the moment, are diluted in futuristic thinking.

But transhumanism has started. Genetic engineering, nanotechnology, biotechnology, information technology, cognitive sciences, robotics and especially artificial intelligence are all advancing unstoppably without any further progress. The implantation of nanotechnology in the human body has been successfully tested for the improvement of certain motor and neuronal functions within the central nervous system. Cyborgs (1) are beginning to see the light of day. Increasingly sophisticated cyber limbs, gradually replacing parts of our own bodies, are advancing with the anticipation that they will eventually surpass them in function and efficiency. Emotional pharmacology modulates our moods in increasingly selective ways, such as, for example, the serotonin inhibitors reuptake syndrome (SIRS)

that achieve strong emotionally bonded behaviours in individuals, so far devoid of psychopathology or sociopathology. Similarly, dopaminergic agonists can improve motor skill acquisition and are associated with increased neural plasticity; cholinesterase inhibitors can improve normal performance in neurodegenerative pathologies such as Alzheimer's disease AD (a disease associated with increased hope) and there are associated new drugs in Phase III. Moreover, new non-addictive stimulants, such as atomoxetine, seem to improve general arousal levels in normal subjects.

Finally, the development of new drugs is not only based on the improvement of underlying diseases, but is also aimed directly at groups of subjects without pathologies. The control group begins to be studied.

These psychotropic drugs, such as AMPAkines and modulators of the AMP-response element binding protein (CREB) cause an intracellular succession of continuous events leading to structural neuronal changes related to the acquisition of higher intellectual capacities such as long-term memories (semantic or episodic) or sustained and alternating attention.

But this onset is not valued as such. It is assumed that these are just isolated advances while waiting for someone to name this succession of facts as emerging transhumanism, which would officially entail a paradigm shift in the human being. And until then we will not begin to assess strictly and objectively all the changes that have occurred, are occurring and will continue to occur in both cognition and emotional management. These changes have already begun, both physiological with an emotional impact, and emotional with an impact on physiology, as epigenetics is discovering day by day thanks to neuroimaging tests.

The emotional neuroscientific community must not abstract itself from a reality that is silently advancing unstoppably, and it must advance in parallel with the scientific achievements that continue to be published in isolation but with a global transhumanist approach.

The only way to achieve this is to use the same scientific method used until now, but with the courage to look ahead without complexes. Observation, detection of needs, elaboration of hypotheses, verification/refutation, interpretation and implementation for continuity. Only in this way will we be able to elaborate criteria that objectify this present and future reality while preserving the general ethical principles that should govern this new posthumanist society.

SCIENCE, CONSCIOUSNESS AND TRANSHUMANISM

There is nothing more provisional than scientific knowledge (Tirapu, 2008). The answers to hypotheses that we ourselves generate come from questions that we ask ourselves based on the observation of the world around us, or even in the field of ethics, philosophy and psychology, from the observation of our own inner world, as well as its interactions.

Scientific answers are replaced by others that refute the first ones on the basis of experimental or statistical tests and that, in turn, expand new questions for further studies. If we move to 110,000 years ago, evolution frames a new paradigm based on the advance of technology, homo sapiens; its first moral, ethical and philosophical conflicts seek to make room for a new knowledge that begins to develop based on new discoveries put at the service of their species, parallel to the development of their capacity for metacognition (González, 2009). However, this parallel development of the internal process together with the advance of technology is a continuum, and if we do not frame both a past and future perspective we will not be able to spontaneously reflect and study the changes in ethical needs, emotional conflicts and repercussions at a global level as social beings tending towards inner technification (Carcar, 2019).

The debates amplified by fiction based on future projections that draw the struggle between machine and human, after the awareness of a "self" of the machines, retreat in the face of currently more pragmatic issues in which the human being itself tends to become robotised by diluting its condition of homo sapiens sapiens evolving towards being the technologically perfected; The H+ (Carvalko, 2012). In this way, human being and machine merge into a future entity in which new ethical conflicts and neuroses will appear, characteristic of this symbiotic evolution, which can be reduced in the same way by the pharmaco-modulatory potential.

Thus, we are faced with a root cause of problems of conscience and thus emotional well-being, which in turn contains the remedy. But to what extent can science have emotion-modulating mechanisms without taking away individual freedom, even from suffering?
The adaptive function of emotions has been proven since the emergence of human beings. Fear leads to pre-reflection, anxiety to the secretion of dopamine, cortisol and glucagon, which allow us to act more efficiently, and anger allows us to defend ourselves against potential enemies. Today's society, on the other hand, has changed and sadness is transformed into depression due to the impossibility of avoiding external stimuli, anxiety comes from continuous stress which leads to a secondary reactive impulse after continuous exposure to the stressful stimulus and anger contains in the same moment the three previous elements mentioned above. Certain emotions have become maladaptive.

As we advance in the necessary transhumanist processes, both emotions and existential conflicts will grow under the neurotic principle of the difference between the consciousness of the present self and the need for the future self. Transhumanism is not a breakthrough with an end but a succession of "improvements" that will entail a succession of expectations.

Therefore, since the appearance of individual consciousness, the progressive rationalisation of our own limits together with that of our own fears has led us to the evolution we know so far. The desire to live longer and better does not entail an end of expectation as we progress in it, but it does contain a discouragement because of the dehumanisation it entails.

This discouragement is based on two concepts:

A) One type of advancement that brings about an improvement in the general human condition is not accessible to every human being on the planet because of the Fragmentation Theorem.

B) Another type of advance that causes an improvement in one part of the human condition, inevitably causes a regression in the rest, since our consciousness works in relative terms, with greater or lesser personal awareness of the relativity of the thought-emotion process.

THE FRAGMENTATION THEOREM

We can define ethics as the discipline that studies good and evil in relation to morality and social behaviour. In other words, it is the moral customs and laws that socially value human behaviour in a community.

These customs and moral laws do not necessarily lead to behaviour in accordance with the former, provoking internal conflicts that are more or less existential and more or less adaptive depending on the consequences of this contradiction on a practical level in the author of each behaviour and in the consequences that fall on his or her environment (excluding specific psychopathies and certain autistic spectrum disorders that prevent the normalised valuation of social cognition and empathy).

According to this premise, the creation of prior ethical norms is of no use for fruitful progress on an emotional level. For this, techno-evolution itself will leave a fragmentation divided into several levels that will need to be recomposed once they appear.

1. Competential posthumanism: Competential conditioning factors will lead to an initial fragmentation of the technologically rich and the technologically poor, with the former being able to access these advances and aggravating the separation between the two groups. Just as in history there was the slave-amo and vassal-lord binomial, the competitive society itself establishes a natural selection in which a few gain access to techno-resources until, by means of social revolution, these resources are forcibly opened up to the rest of the population. The difference with past class struggles is that there is no primary and superior group consciousness, unlike in the past when the mission of some was to keep the existing resources for the few as long as possible while the others fought against the established elite.

   Nowadays, the economic and social system depersonalises the access gap, giving only the most economically capable access to techno-resources without giving those who are not capable a target for the struggle. A non-reductionist example of the essential nature of its consequences would be immunotherapies aimed at improving the life expectancy and quality of life of patients with secondary, metastatic or idiopathic tumours.

   The tried and tested solutions of equalising the social system to the last consequences in order to provide the same resources to the odd and even have had consequences already written down based on the reduction of general resources, in which everything is shared out but in smaller quantities, as well as continuing to maintain the differences between the political elite - much less in proportion - those who direct the rest of the population and therefore reduce individual freedom. As a consequence, two assumptions are made. A) the death without birth of the new technology by neglecting the freedom that allows minimal development of access to experimentation B) the over-attention to technological advancement of a dystopia that neglects the basic needs of the population.

   Thus, we can consider intrinsic interspecies as well as intraspecies competition as a necessary basis for technological progress, being this competition in its two aspects at the same time the very germ of ethical, psychological and social conflict. The advance of transhumanism will intrinsically generate greater differences according to the fragmentation theorem and will provoke an ethical conflict for those who have access to it and a social conflict for those who do not.

2. Delayed transhumanist aggregation: This fragmentation is not a dichotomous value but a cumulative continuum, a fact that disguises the cut-off point of fragmentary consciousness among the new social strata. A few may have access to all techno-resources, a few more may have access to some resources while the majority may have access to a few.

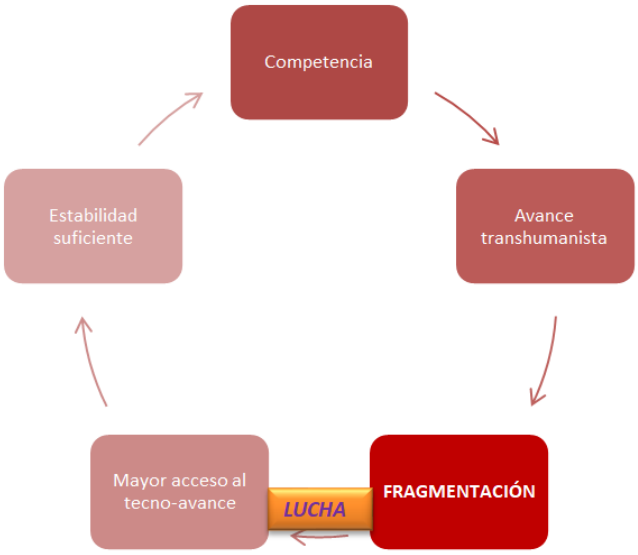
   As the social fragmentation deriving from access to techno-resources, which in turn derives from the techno-advances themselves, increases, an ethical consciousness will be formed that seeks to equalise the masses in terms of general access to them.

The process can be segmented into the following phases: Competition necessary for the emergence of new developments. Access of an economic and power elite to these advances with the consequent progressive emergence of new social techno-classes. Fragmentation of society. Class struggle. Less

limited access to progress as a result of social conflicts in proportion to the level of progress. Widespread stability of progress and social stability. New process of competition

The conclusion that can be drawn today is: on the one hand, this inter-human competition is necessary for the advancement of transhumanism. This competition contains the seeds of fragmentation. The arrival of a stable posthumanism will not be without great social differences, initially silent, causing the progressive appearance of first-quality human beings and second-quality human beings. The struggle for access to techno-quality by the lower classes will lead to suffering, less inequality and stability for the generation of new competition that will generate new advances.

Figure 1. The Fragmentation Theorem.



Source: self-elaboration (2021)

TRANSHUMANISM AND NEUROSIS

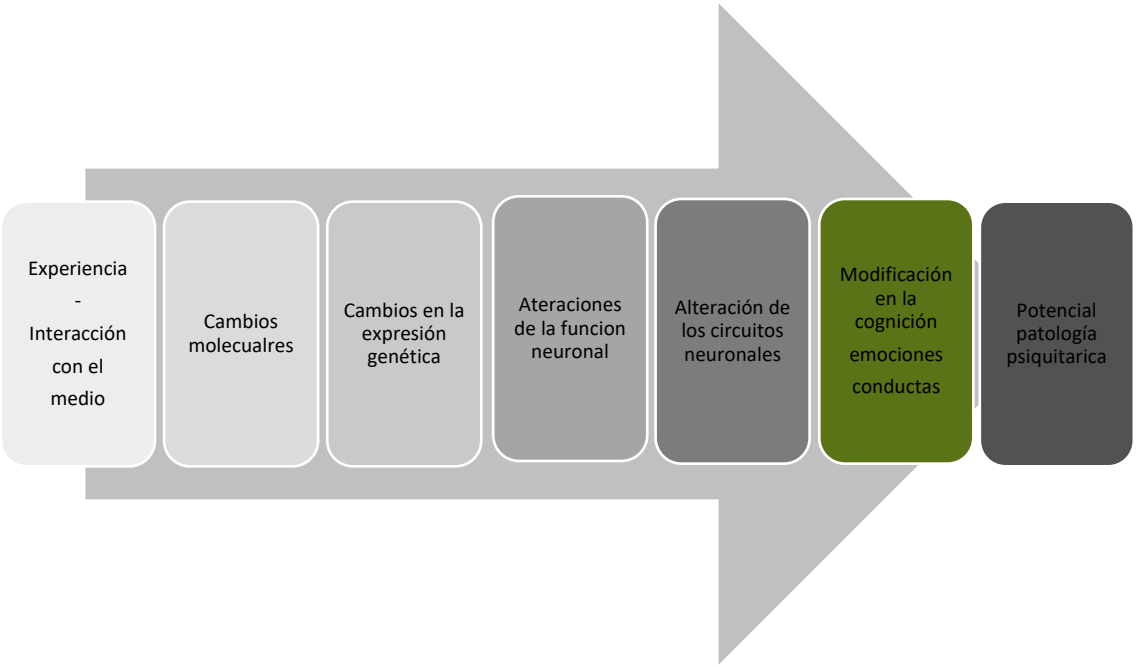
As mentioned in the previous lines, the foundations of neurosis from a physiological point of view have, on the one hand, a genetic component made up of what we call a personality profile that entails certain tendencies apprehended through previous learning that does nothing more than activate these underlying emotions. Just as there is a genetic predisposition for the development of tumours in certain individuals with adequate but minimal environmental activation, while other individuals will not develop any tumours even with greater exposure, there is a predisposition to psycho-emotional conflicts depending on the personality profile that is maintained in the genetic code throughout our lives.

On the other hand, there is a modification of our perceptions in terms of our interaction with the environment in which we live and which, as an imprint that remains, moulds our personality and, therefore, our way of facing each reality.

Why do monozygotic twins, formed from a single zygote and therefore with the same genetic information, present different characters, especially if they have been raised in different environments, why, of the billions of cells that make up our body, even though they have the same genetic information, do each type of cell express some genes and not others, or why do genes express themselves differently depending on our lifestyle? To answer many of these questions is the new field of Genetics called Epigenetics.

Epigenetics is dedicated to understanding the cause and origin of the effects external to genetics that are capable of modifying the expression of genes without altering the nucleotide sequence, i.e. the DNA, and which are transmitted to subsequent generations.

Figure 2. Experiential psycho-epigenetic sequence.



Source: self-elaboration (2021).

The term Epigenetics was coined by Waddington as early as 1942. Today we know that there are three epigenetic mechanisms capable of answering these questions:

1. Chemical modifications of DNA (methylation).
2. Modifications of the histone proteins that are closely linked to it.
3. Non-coding RNAs (ncRNAs).

Such modifications are called epigenetic marks, constituting an additional layer of information called the epigenome.

DNA methylation is essential for the normal development of life, but it can also be the cause of many diseases, such as cancer, premature ageing, dementia, Alzheimer's disease, etc. Fortunately, DNA methylation is a reversible process, which makes it possible to obtain treatments to combat them.

Until not so long ago, knowledge of the hereditary transmission of diseases was limited to the Mendelian model of inheritance, according to which the disorder is due to the mutation of a single gene, either dominant, recessive or X-linked. Later, it was proved that many diseases -as it is the case of psychiatric disorders- respond to a multigenetic model in which several, if not multiple, alleles are involved, each of them contributing in a small proportion to the aetiology of the disorder. This characteristic, present in autism, schizophrenia, bipolar disorder or ADHD, confers a particular complexity to the study of the aetiology, having to point out that most psychiatric disorders are characterised by a high hereditary load, ranging from 0.81 in schizophrenia to 0.37 in depression and

0.50 in conduct disorders (Table 1). On the other hand, the discovery of non-genetic inheritance, which consists in the transmission to the next generation, has led to the development of a new approach to the study of psychiatric disorders.

Table 1. Hereditary burdens in psychiatric disorders.

TRASTORNO	HERENCIA
ESQUIZOFRENIA	0,81
TRASTORNO DEL ESPECTRO AUTISTA	0,80
TRASTORNO BIPOLAR	0,75
DEPRESIÓN MAYOR	0,37
TDAH	0,75
ENFERMEDAD DE ALZHEIMER	0,58

Source: self-elaboration (2021)

Depression is a disorder characterised by a complex aetiopathogenesis involving genetic, neurochemical, neuroendocrine, immunological, environmental and stress resistance factors. The role of epigenetic mechanisms was first suspected when it was observed that treatment with antidepressants has to be long-lasting and sustained over time to be effective and to benefit patients. This observation led to the idea that stable molecular changes occur in depression, affecting brain structures such as the hippocampus, the nucleus accumbens and the prefrontal cortex, which may be due to epigenetic modifications on which the antidepressants in turn act. Depression is characterised by a relatively low heritable burden, much lower than in other psychiatric illnesses, which underlines the importance of environmental factors and other circumstances in aetiopathogenesis. In addition to these heritability differences between monozygotic twins, epigenetic differences have also been observed.

It seems clear that genetic research on depression is one of the most interesting fields in psychiatry today and it is very likely that therapeutic and preventive measures in the coming years will be based on these findings. The progressive study of epigenetic mechanisms is essential to understand how an individual's adaptation to stress results in stable gene expression and appropriate behavioural changes, whereas maladaptation results in gene silencing and increased vulnerability to depression, a phenomenon that can persist throughout life and be passed on to future generations. New research into the epigenetic mechanisms of depression opens up new avenues for prevention and treatment, as these epigenetic alterations, which may be reversible, may become therapeutic targets for new treatments.

Both genetic and environmental factors are involved in the development of schizophrenia. However, there is some controversy about the extent to which the latter act, or when they may be most crucial. A recent study published in Nature Neuroscience has shed light on this question by finding that certain regions of the genome that are epigenetically modified during the early stages of brain development are linked to schizophrenia.

Epigenetic mechanisms may act as a bridge or connection between the environment and genes. As an epigenetic modification, DNA methylation – a modification of the chemical structure of the hereditary material that does not alter its sequence or code, but the way it is interpreted – is a reversible process, as well as potentially sensitive to some environmental factors.

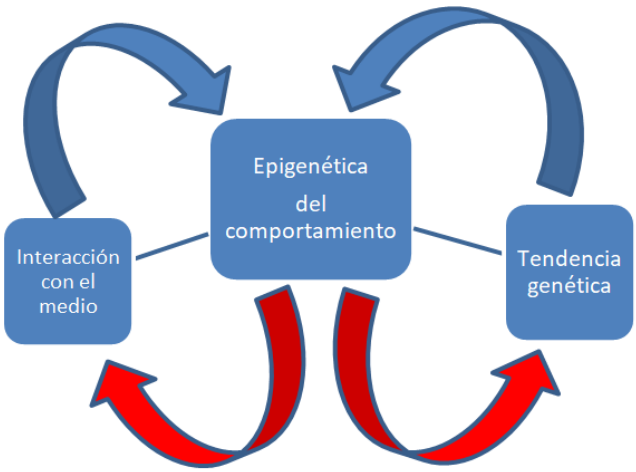
The regulation of gene expression, when, how and where genes are expressed, must be especially precise during brain formation. For this reason, any alterations caused by changes in epigenetic

mechanisms could be critical for the onset of developmental disorders. This is the basis of one of the main hypotheses used to explain the origin of schizophrenia, the hypothesis that the origin of the disease occurred during neurodevelopment.

Genes, therefore, are expressed throughout an individual's life and do so according to genetic programming and the personal and social circumstances in which that life takes place and with which the genes interact, which explains the enormous differences between individuals and between species. This explains the uniqueness of each person's life. Epigenetic mechanisms regulate gene expression and therefore the study of genetics is inseparable from the study of epigenetics and vice versa. Epigenetics is a relatively new field, as are diseases of complex aetiology, and has required technological advances in recent years. Its applications to the understanding of diseases and to the personal reality of patients will come gradually.

Epigenetics, therefore, establishes the link between the constructivist tabula rasa of experience as the only psychomarker and the more Jungian archetypal theory in which we will irrepressibly tend to think, act and feel with a previously defined code.

Figura 3. Psycho-epigenetic feedback.



Source: self-elaboration (2021)

As for the influence of epigenetics on the more organic physiology of the human psyche, the genetic endowment with which the individual is born interacts with the environment in which he or she lives, giving rise to the phenotype, i.e. personal characteristics and diseases. The phenotype conditions personal behaviour and the way of relating to the environment. The environment in turn leads to changes in the genes, either through mutations or epigenetic modifications. This is how the life cycle unfolds, with its numerous changes in phenotype, including possible illness and misfortune, but also progressive personal enrichment and happiness.

ETHICS AND THE POSTHUMAN PSYCHE

One of the greatest dangers facing an ever-advancing biotechnological society is the creation of a two-speed humanity in which the rich advance their organism and the poor cannot, as in the film Elysium: in that fiction the advanced beings abandon the untuned on a troubled and violent Earth to migrate to a paradisiacal satellite. In reality, it is not only the rich who would be upgraded, but people who are

open-minded about it. There would also be technology-averse rich people and not-so-rich people willing to do anything.

On the other hand, in every technological revolution there has been job destruction, conflict, war, violence, class struggle. Now we are facing one of those crises posed by the automation we are already experiencing. The birth of this Fourth Industrial Revolution seems to be gentler than those of the past, but we are already beginning to see casualties as techno-enthusiasm also has conservative overtones. In this digitally reinvented capitalism, perhaps not all human issues will be solved by technology and social aspects will have to be taken into account. And also, its effects elsewhere outside the West, where concepts of the individual or democracy are different. The humanities should play a guiding role in these processes.

But the question is: how to found this new ethic? Where do we start from? On globally shared basic principles, inspired by the perennial wisdom of the worldviews of humanity's diverse religious and spiritual traditions and the various contributions of secular humanist currents of thought," he explains. Despite all the emerging and disruptive technologies that are appearing in the feverish contemporary world, for the author "the essential thing will be to cultivate our interiority, to connect our brain with our heart".

Paradoxically for humanism, and sequentially for transhumanism, it is the circumstance that, thanks to technology, people with a priori disabilities, such as the runner Jacob Pistorius, have become more advanced than human. With his state-of-the-art prosthetic legs, Pistorius ran better than he had ever run with his traditional organic legs.

Although the picture may seem dystopian at times, we must create an ethical vision from the epigenetic basis of philosophy. On the one hand we must maintain the basic general principles of coexistence as a species, including spirituality (religion if it contains a registered mark of God) that allows us to continue not only with our instincts of survival and reproduction but with the development of an inner world capable of connecting with the rest of the people we influence and influence while maintaining and, why not, increasing our empathy for physically unreachable environments. On the other hand, we must look outwards and, with the principles of humanistic science, observe, analyse and create new moral principles adapted to the society that silently calls for a commercially globalised but culturally and sociologically too atomised world. This richness of diversities must be maintained in essence based on respect but, at the same time, we must generate new global sociological responses to maintain a psyche in harmony with our changing environment and our inner conflicts.

Finally, we must maintain and insist on the need for existential optimism despite adversity and hope in the new biotechnological society, as the fruit of a breakthrough we have chosen because we always have a choice. Think about it.

ETHICAL BASES OF TRANSITION

Each transitional base solidifies as new bases are created which, without superimposing themselves on the previous ones, advance parallel to a world in continuous change but with the posthumanist horizon as a point of no return.

Parallel to both technological and eugenic advances are our ethics as individual beings and as social beings with self-awareness, capacity for planning and anticipation of results. This capacity for anticipation, together with the emotion of fear, paralyses any attempt to advance without evaluating the negative conditioning factors. Whether this is a good or bad thing is not the subject of this study and only the reader will have to assess it.

But transhumanist biotechnology has no fear, as its bases are based on unique advances that add up to a whole progression. Advancement and quantitative improvement of its parts and human qualitative improvement is the objective, but how far? how far? how long? This study, after a framing lacking in direct internal reflections, allows itself to leave the following question:

If we could choose how far to go, what would you choose? Perhaps immortality or limitless affectivity. Should we choose between one or the other? Where is the end point?

These questions, initially considered to be of pseudo-philosophical value or lacking in pragmatism, are essentially important, not so much to mark the point and end of this transition but to seal the starting point of the potentiality of these ethical foundations.

As an initial frame of reference, we can sequence some of the proposed objectives of the Transhumanist Party, selecting those considered most relevant to this paper:

1. Support greater longevity through science and technology with a law.
2. Disseminate a pro-science culture with an emphasis on reason and secular values.
3. Provide free education at all levels.
4. Promote morphological freedom; do with one's own body what one wants as long as it does not harm others.
5. Decrease the incarcerated population in the United States by using innovative technology to track crimes outside of prison.
6. Put special emphasis on green technological solutions to make the planet healthier.
7. Support and develop a universal basic income.
8. Restart the space programme with an increase in government resources.
9. Create an international consortium to hold the transhumanist Olympic Games.
10. Work to use science and technology to eliminate all functional diversities (disabilities) in humans.
11. Create an educational and industrial complex in the United States instead of a military industrial complex.
12. Dedicate money from wars against cancer, heart disease, diabetes, not wars in distant countries.

On the other hand, and as an opposite reference, the fundamental goal of transhumanism according to Istvan's fictional book (2014), can be divided into three points:

1. achieving the immortality of human beings.
2. Achieving the omnipotence of human beings
3. Contribute to the cause whose objective is specified in the previous two points.

Without going too deeply into the analysis, several similarities with traditional religions can be observed, with the difference that one does not have to die in order to achieve immortality. In any case, the questions we should ask ourselves in order to understand and help us understand that the ethical foundations are not only opportune but necessary, are to look at the past in terms of our

mistakes for not foreseeing the current advances that have been established, to look at the present to improve our current social system, a reflection of what is to come, and to look to the future in an objective, optimistic way, without mental burdens. It is only a matter of time.

CONCLUSIONS

The teaching of the humanities cannot ignore the new techno-scientific paradigm that is transforming the key concepts of the epistemology with which we have been teaching for several decades. In posthumanism and transhumanism, as we have seen, there is a certain fetishism of the body in its different components, within which the life-giving qualities of the body are produced in malleable and mechanised parts. We are witnessing atomised bodies, where the brain is disembodied and biological components attempt to be digitally reproduced. The human subject is reduced to its materiality, to the functioning of its parts, and its existence to an eminently bioinformatic, biotechnological concept. Today, humanist reflection on the concepts of freedom, autonomy, human nature, among many others, which underpinned the humanisms of the 20th century for so long, is more urgent than ever. We need a critique that runs parallel to scientific and technological development; we need to build a critical posthumanism that, without denying the advances of science, knows how to unveil the rhetorical traps of new racist, sexist and slave-like discourses. This techno-scientific paradigm has opened new paths for different disciplines and theoretical positions, new concepts have entered feminist and post-colonial debates and, in general, human and social sciences. It is urgent to revisit the old questions within the new social and scientific context: when does human life begin and end? who has the power to determine that origin or end? what is the moral status of a human life? what is a life of dignity? The discussion of these questions and others is a task that the humanities must not postpone.

This reductionism forgets, however, that the brain is infinitely more complex than simple neuronal connections, since it has the capacity for reasoning that is logical and illogical, expected and unexpected, chaotic or ordered, creative or not. The decisions that man makes and executes are not only based on reason or objectivity but also on his personal reality, his context, his culture, his idiosyncrasies, etc. All that defines his personal identity and his human nature. In other words, the attribution of mental phenomena is the responsibility of the individual's background of reasons, beliefs and intentions. It is not possible to reduce a psychic description that arises and makes sense in the mental context to reductionist theories about neuronal interactions or to images in a scanner; it is not clear that mind and brain are the same thing. The transhumanist moral stance does not impose any limitations on action. This concept of personhood would also grant personhood to advanced machines. This form of rationalist reductionism forgets that the individual is not a person because his or her rational capacity manifests itself, but that the latter is possible to manifest itself because the individual is a person in himself or herself. As a consequence of his rationalist concept of personhood, he derives a similar concept of dignity: a quality, a kind of excellence that admits of degrees and applies to entities both within and outside the human realm. For Bostrom, for example, dignity would be a quality in human functions like a virtue or an ideal that can be cultivated, encouraged, admired or promoted, without realising that it is reduced to a mere quality control. But who would then establish this quality parameter? Or in other words, who will set the standards of quality for human life? If a few are chosen for this task on the basis of liberal and utilitarian criteria, one inevitably falls into technocratic nepotism, eugenics and social justice problems. Moreover, Bostrom enters into assertions that contradict traditional moral values: further improvements can reduce our Dignity as a Quality. For example, an increased capacity for empathy or compassion can reduce our composure and serenity, leading to a reduction in Dignity as a quality. Given the above, are we less dignified because we have

more compassion? Dignity is also a virtue, but it is not the only one. Therefore, some loss of Dignity as a quality may be compensated by the gain of other virtues. Insisting that dignity in the modern sense consists in who we are and what we have the potential to be, not in our pedigree or our causal origin. This concept of dignity leads him to speak of lives that are more worthy and therefore more valuable than others: we can favour the future generation by being posthuman rather than human, if posthumans would lead more valuable lives than humans would lead. Contrary to what they advocate, we believe that the dignity of the person does not reside in a mere internal or external valuation. The dignity of the person is actually a matter of innate dignity. It is a fundamental intuition, an intrinsic value, which transcends social and cultural barriers and exists because of the peculiar ontological status of the human person, superior to any other personal reality or valuation (e.g., reasoning or not). The transhumanist concept of dignity risks contradicting three fundamental principles of the Universal Declaration of Human Rights: 1) human dignity is universal, something that all individuals possess just by virtue of being human; 2) human dignity is inherent within human nature and is not dependent on their particular achievements or "excellences"; and 3) human dignity applies equally to all persons, not admitting different degrees of it. Again, if the idea of dignity is equated with the idea of autonomy or quality, all instrumental practices on human beings could be justified. However, the imperfection of human beings and their unsatisfied relationship with reality allows them to have aspirations, to progress, to think, to win or to make mistakes... but above all, it allows them to live and transcend; in other words, to be human.

Technology is therefore no longer seen as the solution to all of humanity's problems. Just like the happiness curve with respect to wealth, which slows down its growth with respect to the increase in per capita income once the threshold necessary to cover basic human needs has been crossed, the advance of technology runs the risk of beginning to turn into dehumanism as we move forward on issues that are alien to the essential development of humanity. Now, what are those points alien to human vitality that are not transcendent to the parallel development of ethical and biotheological virtue that provide human beings with greater capacity, freedom or control without beginning the amputation of their theoretical-practical virtues as a society and as individuals that endow them with humanist meaning?

The question is too complex for a single answer, since it depends on each point of reference in terms of culture, individual and social freedom and expectation. No one but us can give it the necessary shape from our morality and culture.

However, just as history has set clear parameters of social and individual ethical awareness as advances have taken hold in society, we will not be able to shape a transhumanist neuroethics congruent with the present and the future without professionals to make their way.

Our duty as psychosocial analysts is to invest time and events to continue with the ethical reflection on transhumanist postulates within a framework of social coexistence and that allows us to maintain a tolerable emotional balance for the continuity of our natural and continuous evolution as the cognitive, emotional and social beings that we are and that the author hopes we will continue to be.

Dedicado a mi padre, mi mentor de vida

KEYWORDS: transhumanism, neuroethics, transhumanist psychology, moral conflict, ethical technology, evolutionary neurosis, posthumanism.

REFERENCES

Annas, G.J., Andrews, L.B. Isasi, R.M., Isasit, R.M. (2002). Protecting the endangered human: toward an international treaty prohibiting cloning and inheritable alterations. *Law Med.,* 28(2-3): 151-178.

Bell, E. Maxwell, B. McAndrews, MP. Sadikot, A. Racine, E. (2011). *Deep brain stimulation and ethics: perspectives from a multisite qualitative study of Canadian neurosurgical centers*. World Neurosurg; 76(6): 537-547.

Bostrom, N. (2005). A History of Transhumanist Thought*. Journal of Evolution and Technology*, 14(1): 1-25.

Bostrom, N. (2008). Ethical issues in human enhancement. In: Ryberg J, Petersen T, Wolf C, editors. *New waves in applied ethics* (pp. 120-152). Michigan: Palgrave Macmillan.

Carcar, J. (2019). El transhumanismo y los implantes cerebrales basados en las tecnologías de la inteligencia artificial: sus perímetros neuroéticos y jurídicos*. Ius et scientia*, 5(1): 157-189. http://doi.org/10.12795/IETSCIENTIA.2019.i01.07

Carvalko, J. (2012). *The Techno-human Shell-A Jump in the Evolutionary Gap.* Sunbury Press.

Chapman, AR (2010). Inconsistency of human rights approaches to human dignity with transhumanism*. Am J Bioeth*., 10(7): 61-63.

Chatterjee, A. (2006). The promise and predicament of cosmetic neurology*. J Med Ethics*; 32(2): 110-113.

Cohen Kadosh, R., Levy, N., O'Shea, J., Shea, N., Savulescu, J. (2012). The neuroethics of non-invasive brain estimulation. *Curr Biol*., 22(4): R108-111.

Fukuyama, F. (2004). *The world's most dangerous ideas: transhumanism.*

Fukuyama, F. (2004). Transhumanism. *Foreign Policy,* 144: 42-43. Recuperado de: http://foreignpolicy.com/2009/10/23/transhumanism/

González, F.E. (2009). Metacognición y aprendizaje estratégico. *Revista Integra Educativa*, *2*(2), 127-136. Recuperado de: http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S1997-40432009000200005&lng=es&tlng=es.

Gonzalez-Melado, F. (2010*).* Transhumanismo: la ideología que nos viene. *Pax et Emerita*, 6(6): 205-228.

Habermas, J. (2003). *The future of human nature. Cambridge*: Polity Press.

Hughes, J. (2004). *Citizen Cyborg: Why Democratic Societies Must Respond to the Redesigned Human of the Future.* Westview Press.

Jotterand, F. (2010). Human dignity and transhumanism: do anthro-technological devices have moral status? *Bioeth.,* 10(7): 45-52.

Kramer, PD. (1997). *Listening to prozac: the landmark book about antidepressants and the remaking of the self*. New York: Penguin Books.

Llano Cifuentes, A (1977). El hombre y su mundo: la estructura psíquica del hombre*. La filosofía en el budismo* (pp. 39-63). Madrid: Dorcas.

McNamee, M.J. Edwards, S.D. (2006). Transhumanism, medical technology and slippery slopes. *Med Ethics,* 32(9): 513-518.

Millán-Puelles, A (1976). *Sobre el hombre y la sociedad*. Madrid: Rialp.

Overall, C. (2017). Tecnologías de mejoramiento de la vida: el significado de la pertenencia a una categoría social. En N. Bostrom y J. Savulescu (eds.), *Mejoramiento humano* (pp. 339-353). España: Teell Editorial.

Palazzani, L (1993). La fundamentación personalista de la bioética. *Cuad Bioet*., 14(2): 48.

Postigo Solana, E (2009). Transhumanismo y post-humano: principios teóricos e implicaciones bioéticas. *Medicina e Morale,* 2:267-282.

Racine E, Illes J. (2006). *Neuroethical responsibilities*. Neurological Science;33(3):269-277.

Safire, W. (2002). Conference introduction: "Our new Prometheum Gift". *Neuroethics mapping the field conference proceedings.* San Francisco: Danna Press.

Sahakian, B.J., Morein-Zamir, S. (2011). Neuroethical issues in cognitive enhancement. *Psychopharmacol.,* 25(2): 197-204.

Tirapu-Ustarroz J., et al (2008). Modelo de funciones y control ejecutivo. *Revista de Neurologia* 46(12). http://doi.org/10.33588/rn.4612.2008252

Young, S. (2006). *Designer Evolution. A Transhumanist Manifesto*. New York.

STATE PHUBBING FULLY MEDIATES THE RELATIONSHIP BETWEEN STATE FEAR OF MISSING OUT AND TIME SPENT ON SOCIAL MEDIA

Yeslam Al-Saggaf

Charles Sturt University (Australia)

yalsaggaf@csu.edu.au

ABSTRACT

Smartphone users on average spend three hours a day on social media. Research shows that state fear of missing out (state FoMo) is a strong predictor of state pubbing, which is the fleeting reaction in which a smartphone user momentarily engages with their smartphone while he/she is having a face-to-face conversation with another person or persons. Research also shows that smartphone users most frequently used social media apps while phubbing others. However, less is known about the relationships among state FoMo, state phubbing, and Time Spent on Social Media (TSoSM). This study examined the relationships among state FoMo, state phubbing and TSoSM. Multiple regression of data collected from an online survey has revealed that state phubbing predicts TSoSM, when controlling for the effects of all other variables in the model, but more importantly, mediation analysis has revealed that state phubbing fully mediates the effect of state FoMo on TSoSM. Considering the negative effects of state FoMo, state phubbing and TSoSM on social media users, understanding the relationships among these factors paves the way for efforts into helping people change harmful habits.

INTRODUCTION

There are about 3.5 billion social media users worldwide spending on average three hours a day on these platforms[26]. The literature indicates that increased Time Spent on Social Media (TSoSM) is associated with numerous detrimental effects and psychological harms. A recent study has found that longer periods of TSoSM are associated with depression, conduct problems and episodic heavy drinking (Brunborg and Burdzovic Andreas, 2019). Another recent study has found that Adolescents who spend more than three hours a day on social media are at a higher risk of developing mental health problems (Riehm et al., 2019) and while one study (Coyne et al., 2020) has found that TSoSM does not impact mental health, a study by Stronge et al. (2019) has found that TSoSM is weakly related to psychological distress. In addition to these psychological problems, a study by Aalbers et al. (2019) has found that TSoSM is associated with higher levels of interest loss, concentration problems, fatigue, and loneliness. As can be seen, the literature is rich with accounts relating to the psychological impacts of TSoSM. However, there is a paucity of research surrounding the question: what predicts the amount of TSoSM and how the prediction happens? The Uses and Gratifications (U&G) theory has been used in the past to shed light on the motives for using social media. This theory argues that individuals actively seek out media that best fulfills their needs (Hollenbaugh et al., 2020). For example, using the U&G framework, researchers have found that engaging with technology may be motivated by a desire to manage state feelings (Elhai and Contractor, 2018). The Media Displacement theory has been used to explain how TSoSM can take away time that could be otherwise spent with family and friends (Tokunag,a 2016). Using a Media Displacement framework, Hall et al. (2019), for example, found that

[26] https://au.oberlo.com/blog/social-media-marketing-statistics

online interactions do take away time from face-to-face interactions. Addiction to social media as a research problem has attracted the attention of many scholars and there is no shortage of research in this area. Karadağ et al. (2015), for example, noted that social media is one of the addiction elements within the smartphones. But these research directions don't address the question: what triggers people to spend longer periods of time on social media and how this triggering takes place? The aim of this study is to investigate what predicts TSoSM and under what circumstances this prediction occurs. Given the dearth of research in this area, understanding what predicts the amount of TSoSM and how is a significant contribution to the literature.

RELATED WORK

There are about 3.5 billion smartphone users worldwide[27], a figure that is forecast to continue to grow. More than 90% of the smartphone users access social media from their smartphones.[28] Dependency on the smartphone has created a new problem, namely phubbing. Phubbing is a fleeting reaction in which a smartphone user momentarily engages with their smartphone while he/she is having a face-to-face conversation with another person or persons (Ivanova et al., 2020). Phubbing has been found to be associated with a number of negative impacts (Al-Saggaf and O'Donnell, 2019a). The impacts of phubbing have been researched in a broad range of settings. Phubbing has been found to be common in workplaces (Roberts and David 2017) and when employers engage in phubbing behaviour, phubbing has been found to decrease employee engagement (Roberts and David, 2017). Phubbing is also common among intimate partners (Roberts and David, 2016) and in situations where phubbing occurred over a long period of time, phubbing has been found to weaken the bond between intimate partners (Roberts and David, 2016), and among married partners phubbing has resulted in reduced ratings of relationship satisfaction, which in turn increased levels of depression (Wang et al., 2017). Phubbing has also been found to negatively impact conversation quality. Phubbing during conversation has been found to decrease the perceived quality of communication, and overall relationship satisfaction (Chotpitayasunondh and Douglas, 2016), and checking the smartphone during a face-to-face interaction can reduce the sense of emotional connection (Nakamura, 2015). In addition, frequent texting via smartphones has been associated with increased smartphone-related conflicts and lower evaluations of relationship quality (Roberts and David, 2016).

Phubbing and social media are related. A study by (Al-Saggaf 2020) has found that participants most frequently used social media apps while phubbing others. Al-Saggaf's (2020) participants reported being significantly more likely to phub using utility apps (i.e., web browser), navigation apps, finance apps, social networking apps, and weather apps than lifestyle apps, entertainment apps, travel apps, news apps, and music apps. Within social networking, participants reported being significantly more likely to phub using Facebook, Facebook Messenger, and Apple/Android Messages than Whatsapp, Pinterest, LinkedIn, Skype, WeChat, and Tumblr.

A number of variables have been found to predict state phubbing. A study by van Rooij et al. (2018) has found that state fear of missing out (FoMo) is a strong predictor of state phubbing and that to a less extent state boredom also predicted state phubbing but that the effect of state loneliness on state phubbing was not significant. The associations among state FoMo, state phubbing and social media inspired the need to investigate the relationships among state phubbing, and/or the state phubbing predictors, namely state FoMo, state loneliness and state boredom, and TSoSM.

[27] https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

[28] https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

FoMo is the need to be in constant contact with one's network, and the fear of missing out on an event where one's network is having fun (Przybylski et al., 2013). FoMo is correlated with social media use (Blackwell et al., 2017; Przybylski et al., 2013), maladaptive smartphone use, and smartphone addiction (Elhai et al., 2016). A study by Elhai et al. (2016) has found that FoMo is the most related factor to problematic smartphone use. However, FoMo was not found to be associated with overall frequency of smartphone use (Elhai et al. 2016). This suggests that even though state FoMo was found to be a strong predictor of state phubbing, it may not predict TSoSM.

Loneliness is the perception of a deficiency within one's network (Bevinn, 2011), triggered either by a small social network size (Al-Saggaf, Utz and Lin, 2016), or unsatisfying relationships as judged by the individuals' personal expectations (Bevinn, 2011). State loneliness is an experience of loneliness that does not persist (Overland, 1991). The relationship between state loneliness and social media is not fully understood, so it is not possible to predict the outcome of the test of the relationship between state loneliness and TSoSM.

State boredom is a fleeting state of under-stimulation in which an individual lacks interest in their surroundings, and is unable to concentrate (Ng et al., 2015). Boredom has been associated with a number of psychological problems especially addiction (Chen & Leung, 2015). Individuals who scored high on boredom played Candy Crush at a much higher level of intensity (Chen & Leung, 2015). But this study focused on leisure boredom; not state boredom. Avoiding feelings of boredom has been found to be a catalyst for using social networking sites (Sheldon, 2008). However, boredom was not found to be associated with smartphone use frequency (Elhai et al., 2018). This suggests that even though state boredom was found to be a weak predictor of state phubbing, it may not predict TSoSM.

METHOD

This study was part of a larger study that looked into the role of personality in smartphone usage. A total of 325 participants completed the Google Forms survey of the study, the link for which was shared in a number of social media sites including sites like Reddit.com. Of the 325 responses received, 19 responses were excluded because the participants indicated ages below 18. As the ethics approval for this study was only for individuals who were 18 and above, the responses of participants under the age of 18 could not be included in the study. Further, as this component of this study is concerned with time spent on social media, the responses of participants who indicated at the time of the study that they did not use any of the social media apps (41 responses) were also excluded from the analysis. In addition, five outliers were detected and unselected bringing the total number of responses used in the analysis to N = 260.

Of the 260 individuals who participated in the study, 23.1% (N=60) of the participants were male and 76.9% (N=200) were female. Participants' ages ranged from 18 to 65, with a mean age of 26.55 (SD = 10.508). Participants came from several countries, including Asian countries, but 40.9% (N=106) of the participants resided in the United States, 25% (N=65) of the participants resided in the United Kingdom and 20.8% (N=54) lived in Australia, with the remaining participants coming from other Asian and Western countries. In terms of the respondents' geographic locations, 47.7% (N = 124) lived in a Metropolitan area and 52.3% (N=136) lived in a Regional area. In terms of social media use, 34.2% (N=89) indicated that they most frequently used Facebook via their smartphone, 36.5% (N=95) said that they used Instagram most frequently, 19.2% (N=50) said that they used Snapchat most frequently and only 10% (N=26) reported that they used Twitter most frequently.

Time spent on social media was measured in minutes per day. Participants were asked: How many minutes, in total, have you spent on social media today? State phubbing was assessed using the state

phubbing scale (Al-Saggaf and O'Donnell, 2019b). The state phubbing scale contained four items rated on a scale from 1 (strongly disagree) to 5 (strongly agree). State boredom was measured using the shortened version of the state boredom scale (Ng et al., 2015). The scale consisted of 19 items rated on a scale from 1 (strongly agree) to 7 (strongly disagree). State fear of missing out was measured using the state fear of missing out scale (Wegmann et al., 2017). The scale consisted of seven items rated on a scale from 1 (not at all true of me) to 5 (extremely true of me). State loneliness was measured using the state loneliness scale (Overland, 1991). The scale consisted of one item ("how often do you feel lonely"), which was responded to on a scale from 1 (Have not yet experienced loneliness) to 6 (Have always felt lonely).

Multiple regression analysis was run in SPSS Version 25 and the mediation analysis was run using Hayes Process Macro for SPSS[29]. The dependent variable was checked for skewness and kurtosis. Excluding the five outliers, mentioned above, improved the conformability of the variable to the assumptions of multiple regression. The variables used to build the model were also assessed for multicollinearity. No problems were found. With regards to the mediation analysis, while all its criteria were met, two of the mediation assumptions, namely normality and linearity were not met. However, given this study is exploratory and because a large sample size and bootstrapping method were used and all the mediation criteria were met –see Figure 2 below, it was judged the effect of not meeting these two assumptions will be small and therefore the mediation analysis continued.

RESULTS

Pearson's correlation was used to assess the relationships between the dependent variable TSoSM and a number of potential predictors; specifically, state phubbing, state FoMo, state boredom, state loneliness, age, gender and geographic location. The correlation analysis has revealed that TSoSM correlated only with state phubbing, state FoMo, age and geographic location. The relationships between TSoSM and state boredom, TSoSM and state loneliness, and TSoSM and gender were not significant and for this reason they were not included in the regression model. For the Pearson's correlations among the variables in the regression model see Table 1 below. The means and standard deviations for these variables are listed in Table 2 below. As can be seen from Table 2, on average participants spent approximately 40 minutes on social media on the day they completed survey. The mean of 2.213 for state FoMo suggests that overall the participants were neither experiencing higher levels of state FoMo, nor lower levels of state FoMo; rather they were almost in the middle. In the case of state phubbing a mean of 3.269 suggests that participants likelihood to engage in state phubbing was at a position between 'neutral' and 'agree'; that is, neither between 'agree' and 'strongly agree' nor between 'disagree' and 'neutral.'

Table 1. Pearson's Correlations among Variables.

	Variable	1	2	3	4	5
1	TSoSM	-	.278**	-.132*	-.181**	.343**
2	State FoMo		-	-0.083	-.220**	.549**
3	Geographic Location			-	.221**	-.144*
4	Age				-	-.278**
5	State phubbing					-

** Correlation is significant at the 0.01 level

** Correlation is significant at the 0.05 level*

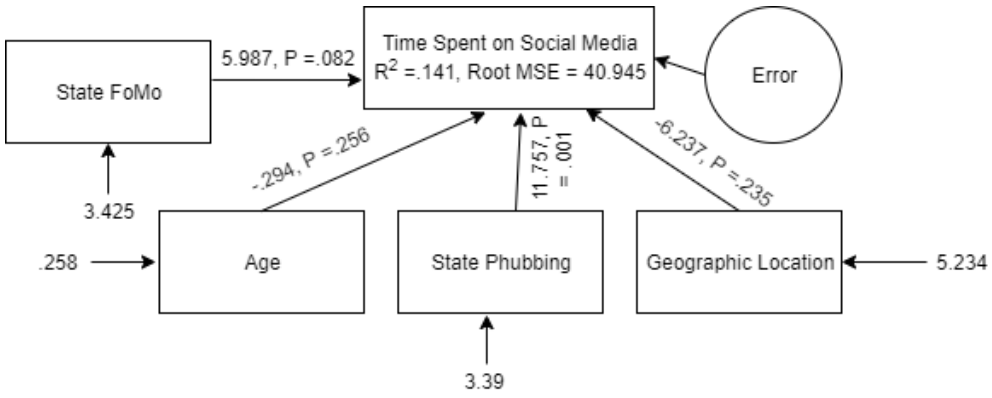
[29] http://processmacro.org/index.html

Table 2. Means and Standard Deviations of Variables.

Variable	Mean	Std. Deviation	N
TSoSM	40	43.8	260
State FoMo	2.213	0.9	260
Geographic Location	1.52	0.5	260
Age	26.55	10.508	260
State phubbing	3.269	.918	260

The unstandardized regression coefficients are measures of effect size. The multiple regression analysis has revealed that state phubbing is a strong predictor of TSoSM. For a unit change in state phubbing, there are 11.757 units change in TSoSM while controlling for the effects of all other independent variables in the model. Considering TSoSM was measured in minutes per day, each additional unit of change in state phubbing leads to 11.757 minutes of additional time spent on social media. Figure 1 below shows the results of the regression model along with the regression coefficients, their associated standard errors, and their relevant p-values.

Figure 1. Results of the regression model along with the regression coefficients, their associated standard errors, and their relevant p-values.



While the effects of state FoMo, geographic location and age were not statistically significant, statistical insignificance of individual predictor variables within a statistically significant regression model does not suggest these individual predictor variables should not be considered when reporting the results. Since the regression model is statistically significant, these individual predictor variables should be counted too. The unstandardized regression coefficient of state FoMo is positive and indicates that an increase of a unit change in state FoMo leads to an increase of almost six minutes of time spent on social media. The unstandardized regression coefficient of geographic location indicates that participants in regional areas spend 6.237 minutes less on social media compared to their metropolitan counterparts. Similarly, the unstandardized regression coefficient of age indicates that as age increases, TSoSM decreases by 30 seconds.

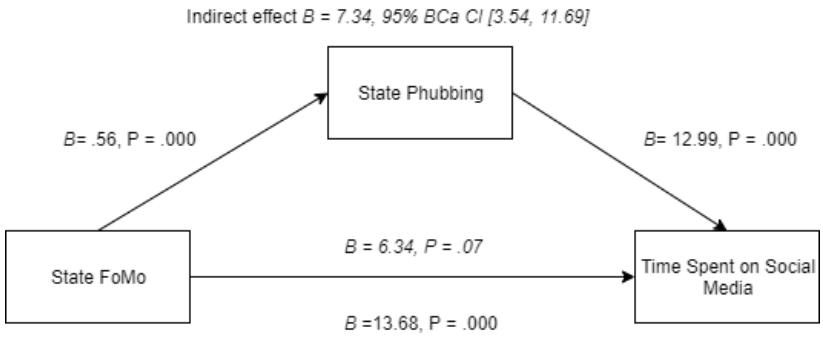
The regression analysis has generated an R^2 of 0.141. The R^2 value of 0.141 shows that the model explained 14.1% of the variation of the dependent variable, TSoSM, around its mean. This is a significant percentage. However, the large value of the Root MSE (SEE) (41.8), indicates that there are issues with the goodness of the fit of the model or more importantly that other variables not included in the model are playing a bigger role in predicting TSoSM. Further research can reveal these variables.

Regression assess the effect of an independent variable on the dependent variable while controlling for the effects of all other independent variables in the model. Mediation address the question how the independent variable predicts the dependent variable. That is, it explains the mechanism of the relationship between the independent variable and the dependent variable. The mediation analysis was conducted using Hayes Process macro for SPSS[30]. Percentile bootsrapped 95% confidence interval were calculated based on 5000 samples.

The Total Effect, measured without including the mediator in the model, shows that state FoMo significantly predicted TSoSM, B = 13.68, t = 4.65, p = .000. The R^2 value of 0.08 shows that the model explains 8% of the variance in TSoSM. State FoMo has a positive relationship with TSoSM, as shown in Figure 2.

The Direct Effect, which includes the mediator variable in the model, was not significant. This is one of the conditions for the occurrence of the mediation and Figure 2 shows this condition is met. This suggests that state FoMo does not directly predict TSoSM when the mediator is included. The indirect effect of state FoMo on TSoSM through state phubbing was significant, B = 7.34, 95% BCa CI [3.54, 11.69]. The R^2 value indicates that the model explains 13% of the variance in TSoSM. Given the direct effect was not significant and the indirect effect was significant, state phubbing fully mediates the relationship between state FoMo and TSoSM, as shown in Figure 2. This means that the entire effect of state FoMo on TSoSM is transmitted through state phubbing.

Figure 2. Results of the mediation analysis.



DISCUSSION AND CONCLUSION

The relationship between state FoMo and Time Spent on Social Media (TSoSM) is not fully understood. Previous research has found that state FoMo is a strong predictor of state phubbing (van Rooij et al. 2018) and the current study has confirmed that state phubbing is correlated with TSoSM. The study then examined the relationships among state phubbing, state FoMo and TSoSM. The regression analysis has revealed that state phubbing predicts TSoSM. The regression analysis has shown that each additional unit of change in state phubbing leads to 11.757 minutes of additional time spent on social media. This indicates that the momentary reaction of checking the smartphone during a face-to-face conversation, increases the TSoSM. The fleeting feelings of boredom and loneliness were not correlated with TSoSM. Therefore, these variables were not included in the regression model. The same goes for the demographic variable gender. While state loneliness and gender did not predict state phubbing in Al-Saggaf and O'Donnell's (2019a) study, which is why they were not expected to predict TSoSM, it is not clear why state boredom, which had a weak effect on state phubbing (Al-Saggaf

[30] http://processmacro.org/index.html

and O'Donnell 2019b), did not play any role in predicting TSoSM. This is a question for future research, but the result is not unexpected either as boredom was not found to be associated with smartphone use frequency (Elhai et al. 2018). Age had a significant negative effect on TSoSM, suggesting as age increase, TSoSM decreases. This finding is consistent with previous research including Brunborg and Burdzovic Andreas (2019) study. With regards to the demographic variable geographic location, the regression analysis has shown that participants in regional areas spent less time on social media, 6.237 minutes less, compared to their counterparts in metropolitan areas. This finding is interesting and should be looked into in a future study.

The study then tested whether state phubbing mediates the relationship between state FoMo and TSoSM. The mediation analysis has revealed that state FoMo affects TSoSM via state phubbing. State phubbing fully mediated the effect of state FoMo on TSoSM. The result that state FoMo predicts TSoSM through state phubbing may mean that the fleeting feeling of state FoMo become more intense when smartphone users are having face-to-face conversations with others which is why they momentarily check their social media feeds via their smartphones, i.e. phub their conversationalists. This temporary checking of the social media apps during face-to-face conversations to relief the fleeting feeling of FoMo makes state phubbing a coping mechanism. This conclusion can be explained by the U&G theory (Papacharissi and Rubin 2000), which postulates that people may instantly engage with their smartphones to relieve a state feeling, in this case state FoMo. But as users phub others to overcome state FoMo they end up spending more time on social media. These conclusions are all new and given the negative impacts of phubbing and social media they should be the subject of a future study. Indeed, a future study should empirically, preferably experimentally, tests whether the short-lived checking of social media apps during face-to-face conversations (state phubbing) serves as a coping mechanism to overcome the short-lived feeling of FoMo.

This paper makes a significant contribution to the literature. This study is the first that found state phubbing is a strong predictor of TSoSM and is the first that found that the effect of state FoMo on TSoSM is fully transmitted through state phubbing. State FoMo, state phubbing and TSoSM are associated with detrimental effects and psychological harms. Understanding the relationships among state FoMo, state phubbing and TSoSM paves the way for efforts into helping people change harmful habits. The finding that state phubbing is a mediator between state FoMo and time spent on social media which smartphone users momentarily go to when they are having face-to-face conversations with others to overcome state FoMo is a significant discovery that should be examined experimentally in future research.

KEYWORDS: Fear of missing out, phubbing, smartphone, social media, time, regression, mediation.

REFERENCES

Aalbers, G., McNally, R. J., Heeren, A., de Wit, S., and Fried, E. I. (2019). Social Media and Depression Symptoms: A Network Perspective, Journal of Experimental Psychology: General (148:8), American Psychological Association Inc., pp. 1454-1462. https://doi.org/10.1037/xge0000528

Al-Saggaf, Y. (2020) App Use While Phubbing. In: Ahram T., Taiar R., Gremeaux-Bader V., Aminian K. (eds) Human Interaction, Emerging Technologies and Future Applications II. IHIET 2020. *Advances in Intelligent Systems and Computing*, vol 1152. pp. 238-244. Springer, Cham. https://doi.org/10.1007/978-3-030-44267-5_36

Al-Saggaf, Y. and O'Donnell, S. B. (2019a). Phubbing: Perceptions, reasons behind, predictors, and impacts. *Human Behavior and Emerging Technologies*, 1(2), 132-140.

Al-Saggaf, Y. and O'Donnell, S. B. (2019b). The Role of State Boredom, State of Fear of Missing Out and State Loneliness in State Phubbing. Proceedings of the 30th Australasian Conference on Information Systems, Perth, Australia. December 9th -11th, 2019.

Al-Saggaf, Y., Utz, S. & Lin, R. (2016). Venting negative emotions on Twitter and the number of followers and followees. *International Journal of Sociotechnology and Knowledge Development*. 8(1), 45-56.

Blackwell, D., Leaman, C., Tramposch, R., Osborne, C., and Liss, M. 2017. "Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addiction," Personality and Individual Differences (116), October, pp 69-72. https://doi.org/10.1016/j.paid.2017.04.039

Brunborg, G. S., and Burdzovic Andreas, J. (2019). Increase in Time Spent on Social Media Is Associated with Modest Increase in Depression, Conduct Problems, and Episodic Heavy Drinking, *Journal of Adolescence* (74), Academic Press, pp. 201-209. (https://doi.org/10.1016/j.adolescence.2019.06.013

Chen, C., & Leung, L. (2015). Are you addicted to Candy Crush Saga? An exploratory study of linking psychological factors to mobile social game addiction. *Telematics and Informatics*, (May). http://doi.org/10.1016/j.tele.2015.11.005

Chotpitayasunondh, V., and Douglas, K. M. (2016). How 'Phubbing' Becomes the Norm: The Antecedents and Consequences of Snubbing via Smartphone, *Computers in Human Behavior* (63), Elsevier, pp. 9-18. https://doi.org/10.1016/j.chb.2016.05.018

Coyne, S. M., Rogers, A. A., Zurcher, J. D., Stockdale, L., and Booth, M. (2020). Does Time Spent Using Social Media Impact Mental Health?: An Eight Year Longitudinal Study, *Computers in Human Behavior* (104), Elsevier, p. 106160. https://doi.org/10.1016/j.chb.2019.106160

Elhai, J. D., and Contractor, A. A. (2018). Examining Latent Classes of Smartphone Users: Relations with Psychopathology and Problematic Smartphone Use, *Computers in Human Behavior* (82), Elsevier, pp. 159-166. https://doi.org/10.1016/j.chb.2018.01.010

Elhai, J. D., Levine, J. C., Dvorak, R. D., and Hall, B. J. (2016). Fear of Missing out, Need for Touch, Anxiety and Depression Are Related to Problematic Smartphone Use, *Computers in Human Behavior* (63), Elsevier Ltd, pp. 509-516. https://doi.org/10.1016/j.chb.2016.05.079

Franchina, V., Vanden Abeele, M., Van Rooij, A. J., Lo Coco, G., & De Marez, L. (2018). Fear of missing out as a predictor of problematic social media use and phubbing behavior among Flemish adolescents. International journal of environmental research and public health, 15(10), 2319.

Hall, J. A., Johnson, R. M., and Ross, E. M. (2019). Where Does the Time Go? An Experimental Test of What Social Media Displaces and Displaced Activities' Associations with Affective Well-Being and Quality of Day, *New Media & Society* (21:3), SAGE Publications, pp. 674-692. https://doi.org/10.1177/1461444818804775

Ivanova, A., Gorbaniuk, O., Błachnio, A., Przepiórka, A., Mraka, N., Polishchuk, V., and Gorbaniuk, J. (2020). *Mobile Phone Addiction, Phubbing, and Depression Among Men and Women: A Moderated Mediation Analysis*, pp. 1-14. https://doi.org/10.1007/s11126-020-09723-8

Karadağ, E., Tosuntaş, Ş. B., Erzen, E., Duru, P., Bostan, N., Şahin, B. M., Çulha, İ., and Babadağ, B. (2015). Determinants of Phubbing, Which Is the Sum of Many Virtual Addictions: A Structural Equation Model, *Journal of Behavioral Addictions* (4:2), pp. 60-74. https://doi.org/10.1556/2006.4.2015.005

Nakamura, T. (2015). The Action of Looking at a Mobile Phone Display as Nonverbal Behavior/Communication: A Theoretical Perspective, *Computers in Human Behavior* (43), Elsevier, pp. 68-75. https://doi.org/10.1016/j.chb.2014.10.042

Papacharissi, Z., and Rubin, A. M. (2000). Predictors of Internet Use, *Journal of Broadcasting and Electronic Media* (44:2), Routledge, pp. 175-196. https://doi.org/10.1207/s15506878jobem4402_2

Przybylski, A. K., Murayama, K., DeHaan, C. R., and Gladwell, V. (2013). Motivational, Emotional, and Behavioral Correlates of Fear of Missing Out, *Computers in Human Behavior* (29:4), Pergamon, pp. 1841-1848. https://doi.org/10.1016/J.CHB.2013.02.014

Riehm, K. E., Feder, K. A., Tormohlen, K. N., Crum, R. M., Young, A. S., Green, K. M., Pacek, L. R., La Flair, L. N., and Mojtabai, R. (2019). Associations between Time Spent Using Social Media and Internalizing and Externalizing Problems among US Youth," *JAMA Psychiatry* (76:12), American Medical Association, pp. 1266-1273. https://doi.org/10.1001/jamapsychiatry.2019.2325

Roberts, J. A., and David, M. E. (2016). My Life Has Become a Major Distraction from My Cell Phone: Partner Phubbing and Relationship Satisfaction among Romantic Partners, *Computers in Human Behavior* (54), Elsevier, pp. 134-141. https://doi.org/10.1016/j.chb.2015.07.058

Roberts, J. A., and David, M. E. (2017). Put down Your Phone and Listen to Me: How Boss Phubbing Undermines the Psychological Conditions Necessary for Employee Engagement, *Computers in Human Behavior* (75), Elsevier, pp. 206-217. https://doi.org/10.1016/j.chb.2017.05.021

Sheldon, P. (2008). Student favorite: Facebook and motives for its use. Southwestern Mass *Communication Journal*, 23(2), 39-53.

van Rooij, A. J., Lo Coco, G., De Marez, L., Franchina, V., and Abeele, M. Vanden. (2018). Fear of Missing out as a Predictor of Problematic Social Media Use and Phubbing Behavior among Flemish Adolescents, *International Journal of Environmental Research and Public Health* (15:10). https://doi.org/10.3390/ijerph15102319

Tokunaga, R. S. (2016). An Examination of Functional Difficulties From Internet Use: Media Habit and Displacement Theory Explanations, *Human Communication Research* (42:3), Blackwell Publishing, pp. 339-370. https://doi.org/10.1111/hcre.12081

Wang, X., Xie, X., Wang, Y., Wang, P., and Lei, L. (2017). Partner Phubbing and Depression among Married Chinese Adults: The Roles of Relationship Satisfaction and Relationship Length, *Personality and Individual Differences* (110), Elsevier, pp. 12-17. https://doi.org/10.1016/j.paid.2017.01.014

Wegmann, E., Oberst, U., Stodt, B., and Brand, M. (2017). Online-Specific Fear of Missing out and Internet-Use Expectancies Contribute to Symptoms of Internet-Communication Disorder, *Addictive Behaviors Reports* (5), Elsevier, pp. 33-42. https://doi.org/10.1016/j.abrep.2017.04.001

ETHICAL DIGITAL COMUNICATION. INFLUENCE OF FEAR ON THE INTENTION TO GET VACCINES FOR COVID-19

Jorge Pelegrín-Borondo, Orlando Lima Rua, Leonor González Menorca, Sandrina Teixeira

Universidad de La Rioja (Spain), CEOS. PP/Polytechnic of Porto/ISCAP (Portugal),
Universidad de La Rioja (Spain), CEOS. PP/Polytechnic of Porto/ISCAP (Portugal)

jorge.pelegrin@unirioja.es; orua@iscap.ipp.pt; leonor.gonzalez@unirioja.es; sandrina@iscap.ipp.pt

INTRODUCTION

The disease produced by the SARS-CoV-2 virus (called Covid-19) was identified in December 2019 in Wuhan (China). Since then, the number of people that died has increased worldwide, reaching more than 3,000,000 people. This situation is also seriously affecting the economy (O'Grady, 2020). Most countries are expected to go through recessions (WHO, 2020), with the consequent suffering of people both from dead loved ones and their financial problems.

We are in a situation in which ethics requires that we worry about how to remedy this situation and one of the solutions is to control effective vaccines against the virus. When this research was carried out, we did not have any proven effective vaccine and we are in a research phase for the development of these vaccines. In this phase, vaccines are tested in groups of healthy people to determine if they are safe and if there are possible side effects (CDC, 2014).

Due to the importance of vaccines to control the Covid-19 pandemic, the factors that influence the acceptance of potential vaccines should be investigated. If a vaccine is ready but a large part of the population refuses to receive it, it will have no beneficial effects. Several international studies have been conducted on the percentage of the population that wishes to receive the Covid-19 vaccine (Malik et al., 2020). Some studies show that the willingness to get vaccinated against Covid-19 is insufficient (Lazarus et al., 2020). So the mere availability of a vaccine will not guarantee its success if a large enough number of people are not vaccinated. Therefore, it is imperative to investigate the factors that influence the willingness to be vaccinated. The acceptance of the vaccine could vary and be based on different factors (Dubé et al., 2018; Fu et al., 2017; Quinn et al., 2019; Sarathchandra et al., 2018).

In this sense, the information that people receive is crucial when it comes to accepting a vaccine. Communications by non-traditional means, such as social networks, have opened a new means of communication. In the acceptance of vaccines, these new ways of information have been used by some people. There are people who, based on fake news, deny the existence of the Covid-19 disease and people who, due to their fear and misinformation, have taken unhealthy actions to avoid getting infected (such as injecting disinfectant).

Thus, one of the essential components in the emotions' formation is the existence of something that generates that emotion (Pelegrín-Borondo et al., 2017). That something can be news about what is being valued (Pelegrín-Borondo et al. 2019). In this way, the news we receive can generate emotions that in turn influence our decisions.

At the same time, the decision to use a vaccine is affected by the emotions that people feel (Manca, 2018) and their emotional states (Luz, Brown, & Struchiner, 2019). Several investigations have shown that the negative emotions that a person feels influence the intention to use a certain vaccine (Chapman & Coups, 2006). Thus, intense negative emotions related to vaccines increase the possibility

of changing a specific behavior of people (Chan, Cheng, Tam, Huang, & Lee, 2015). One of the most studied negative emotions in the acceptance of the vaccine is fear. The fear of virus infection could be used to convince people of the need to be vaccinated (Lau, Yang, Tsui, and Kim, 2003). In this sense, fear can influence in several ways: (1) Fear of the side effects of the vaccine can negatively include the intention to use said vaccine (Abebe et al., 2019; Anraad et al., 2020; Kyaw et al., 2019; Otieno et al., 2020); (2) the fear of needles and bleeding when they give you a vaccine is a reason to refuse to be vaccinated (Luz et al., 2019), and (3) the fear of contracting the disease caused by a virus positively influences the acceptance of a vaccine (Anraad et al., 2020; Nguyen et al., 2020).

Regarding fear of the disease, it has been shown that consumer fear of infection can stimulate acceptance of the vaccine that would prevent that infection (Anraad et al., 2020). Furthermore, the fear of contracting an infection is a positive motivator for the tendency to follow the instructions of the health authorities and accept vaccination (Poland, 2010).

Related to the above, in this research we have communicated to the respondents news received digitally about the stoppage of trials in a vaccine for Covid-19 due to reports of a serious adverse event in a volunteer. Subsequently, we have analyzed how the emotion of fear affects the side effects of vaccines for Covid-19 and the fear of the Covid-19 disease produced by the SARS-CoV-2 virus, the intention to use different vaccines, and the relationship between this fear and the acceptance of different vaccines.

THEORETICAL FRAMEWORK

There is a broad consensus among researchers on the existence of the influence of emotions on how people behave (Russell, 2003) and on the way, they make their decisions about products (e.g., Han et al., 2006; Pelegrín et al., 2019). There is no single definition of emotion. However, there is a certain consensus among researchers that there is a set of common characteristics that all emotions have and that therefore allows defining them under these characteristics (Pelegrín-Borondo et al., 2017; Russell, 2009): there is always a stimulus that produces emotions; emotions feel qualitatively unique; an emotion produces a physiological reaction; emotion will produce a tendency to action; when there is emotion there is a cognitive evaluation (distinguishing itself from visceral reactions in which there is no cognitive evaluation); emotions last a short time (moods last much longer); in the emotions, we can graduate the pleasure or displeasure they produce.

For this research, the cause of the emotion is fundamental, since the cause may be the object that is being evaluated, the evaluation process, or something alien to both the evaluated object and the evaluation process (Pelegrín-Borondo et al., 2017). In addition, the cause of the emotion can be mixed, this is our case since we are going to provide the respondents with news about a problem in a vaccine for covid-19 and then we will analyze how fear of the vaccine and fear of the covid-19 disease in the intention to get vaccinated.

The basic emotions approach has been one of the most used by researchers (Pelegrín-Borondo et al., 2015). In the basic emotions approach, it is considered that people are capable of recognizing emotions and differentiating between different types of emotions (Russell & Barrett, 1999). In this way, each type of emotion brings together a set of emotions that resemble it. Each type is what we call basic emotion (Ortony & Turner, 1990; Scherer, 2005). There are infinite emotions, but basic emotions are like the basic colors that represent the infinite set of colors close to them.

Different lists of basic emotions have been established in the scientific literature. Some of the lists of basic emotions most used in the study of consumer behavior are DES, CES, and PANAS (Pelegrín-

Borondo et al., 2015). Although these are the most widely used scales, there are many more. However, it is difficult to find a scale of basic emotions that does not include fear among them.

If we join these last comments with what has already been said about the trigger for emotions, we can say that the emotion of fear can be triggered by several triggers. Thus, fear can be produced by a news item that deals with the possible effects of vaccines or by our fear of the disease. In this sense, several researchers have shown that the fear of the Covid-19 disease is different than the fear of the side effects of the Covid-19 vaccine (Anraad et al., 2020; Nguyen et al., 2020). In this sense, fear of producing a disease positively affects the intention to get vaccinated (Nguyen et al., 2020; Patil et al., 2020; Anraad et al., 2020), while fear of side effects (a short-term or temporary and long-term or permanent) negatively influence the intention to get vaccinated (Abebe et al., 2019; Borena et al., 2016; Cordoba-Sanchez et al., 2019; Kyaw et al., 2019; Maltezou et al. al., 2019; Nguyen et al., 2020; Otieno et al., 2020).

There is a broad consensus among researchers on the existence of influence of emotions on how people behave (Russell, 2003) and on the way they make their decisions about products (e.g., Han et al., 2006; Pelegrín et al., 2019). There is no single definition of emotion. However, there is a certain consensus among researchers that there is a set of common characteristics that all emotions have and that therefore allows defining them under these characteristics (Pelegrín-Borondo et al., 2017; Russell, 2009): there is always a stimulus that produces emotions; emotions feel qualitatively unique; an emotion produces a physiological reaction; emotion will produce a tendency to action; when there is emotion there is a cognitive evaluation (distinguishing itself from visceral reactions in which there is no cognitive evaluation); emotions last a short time (moods last much longer); in the emotions we can graduate the pleasure or displeasure they produce. For the present research, the cause of the emotion is fundamental, since the cause may be the object that is being evaluated, the evaluation process or something totally alien to both the evaluated object and the evaluation process (Pelegrín-Borondo et al., 2017). In addition, the cause of the emotion can be mixed, this is our case since we are going to provide the respondents with news about a problem in a vaccine for covid-19 and then we will analyze how fear of the vaccine and fear of the covid-19 disease in the intention to get vaccinated. The basic emotions approach has been one of the most used by researchers (Pelegrín-Borondo et al., 2015). In the basic emotions approach, it is considered that people are capable of recognizing emotions and differentiating between different types of emotions (Russell and Barrett, 1999). In this way, each type of emotion brings together a set of emotions that resemble it. Each type is what we call basic emotion (Ortony & Turner, 1990; Scherer, 2005). There really are infinite emotions, but basic emotions are like the basic colors that represent the infinite set of colors close to them. Different lists of basic emotions have been established in the scientific literature. Some of the lists of basic emotions most used in the study of consumer behavior are DES, CES and PANAS (Pelegrín-Borondo et al., 2015). Although these are the most widely used scales, there are many more.

However, it is difficult to find a scale of basic emotions that does not include fear among them. If we join these last comments with what has already been said about the trigger for emotions. We can say that the emotion of fear can be triggered by several triggers. Thus, fear can be produced by a news item that deals with the possible effects of vaccines or by our fear of the disease. In this sense, several researchers have shown that the fear of the Covid-19 disease is different than the fear of the side effects of the Covid-19 vaccine (Anraad et al., 2020; Nguyen et al. 2020). In this sense, fear of producing a disease positively affects the intention to get vaccinated (Nguyen et al., 2020; Patil et al., 2020; Anraad et al., 2020), while fear of side effects (a short-term or temporary and long-term or permanent) negatively influence the intention to get vaccinated (Abebe et al., 2019; Borena et al., 2016; Cordoba-Sanchez et al., 2019; Kyaw et al., 2019; Maltezou et al. al., 2019; Nguyen et al., 2020; Otieno et al., 2020).

METHODOLOGY

A survey was carried out with Spanish residents. To see the influence of a news item that could cause fear on the acceptance of vaccines, before starting with the survey questions, respondents were informed of the following news:

"Trials on the vaccine developed for Covid-19 by the University of Oxford and AstraZeneca were suspended on September 9, 2020, following reports of a 'serious adverse event' in a volunteer.". The objective was that this news could generate fear and later verify if the influence of fear of the vaccine and the virus affects the intention to get vaccinated.

Potential respondents were contacted via telematics, requesting their participation. Gender quotas and three age ranges were established. As the respondents of each rank were obtained, the effort to contact was concentrated in the other ranks. The survey was self-administered and completed online. The information was collected from Tuesday, September 9, 2020 to Wednesday, September 16, 2020. 600 valid surveys were obtained. The characteristics of the sample are: (a) woman 55% and men 45%, and (b) From 17 to 30: 33%; from 31 to 50: 33%; 51 or older: 34%.

To measure fear of Covid-19 disease, the scale of Nguyen et al (2020) was adapted. The intention to use various vaccines was also asked [CanSino (China), SputnikV (Russia), Moderna (USA), AstraZeneca (UK-USA)]. To measure this variable, the item developed by Venkatesh and Davis (2000) on intention to use has been adapted. The measure used was an 11-point Likert-type scale, from no agreement (0 points) to total agreement (10 points).

Structural equation modeling (SEM) has been used to carry out the analysis of the influence of fear of the vaccine and fear of the Covid-19 disease. Specifically, among the various options of SEM techniques, the Consistent Partial Least Square (PLSc) has been used. The rationale for this decision is that traditional Partial Least Squares (PLS) tend to underestimate regression coefficients and tend to skew factor loadings upwards (Gefen et al., 2011). This decision has also been made because "PLSc avoids the excessive amount of Type I and Type II errors that can occur if traditional PLS or regression on sum scores is applied to estimate structural equation models with reflective measurement models" (Dijkstra & Henseler, 2015a, p. 299). Since the present investigation corresponds to this case, it has been decided to use PLSc. On the other hand, the set of PLS techniques are less sensitive to problems with the violation of data normality assumptions (Ram et al., 2014).

RESULTS

Regarding the results, Table 1 shows the arithmetic mean and the standard deviation of the items for fear of the disease. As can be seen, the fear of being infected is high (6.60) and the fear of transmitting Covid-19 is even higher (7.86). Here we observe an interesting result from the ethical point of view: selfishness (fear of getting infected) is less than our fear of hurting other people (fear of transmitting covid-19). Furthermore, we have observed that fear of the temporary effects of the vaccine is less important than fear of the permanent effects of the vaccine.

Table 2 shows the results of the items for the intention to use the different vaccines. This is an also interesting result, the vaccines developed by countries with greater cultural ties with Spain (the United Kingdom and the United States) have greater acceptance than vaccines developed by countries with fewer cultural ties (China and Russia).

Table 1. Arithmetic mean of the items on fear of Covid-19 disease and vaccine.

Items fear of the covid-19	Average
I am afraid of getting Covid-19	6.60
I am afraid of transmitting Covid-19 to others	7.86
I am afraid of the temporary effects of the vaccine	6.75
I am afraid of the permanent effects of the vaccine	7.23

Table 2. Arithmetic mean of vaccine intention.

Intention to use vaccine	Average
I intend to use the **AstraZeneca (UK-EEUU)** vaccine	5.07
I intend to use the **CanSino (China)** vaccine	2.54
I intend to use the **Sputnik (Russia)** vaccine	2.14
I intend to use the **Moderna (EEUU)** vaccine	3.38

In Figures 1, 2, 3, and 4 we observe how fear of contracting Covid-19 and transmitting Covid-19 influence the intention to use the different vaccines. In Figures 5, 6, 7, and 8 we observe how fear of the temporary effects of vaccine and fear of permanent effects of the vaccine influence the intention to use the different vaccines.

Figure 1. Fear of covid-19. AstraZeneca (UK-EEUU).

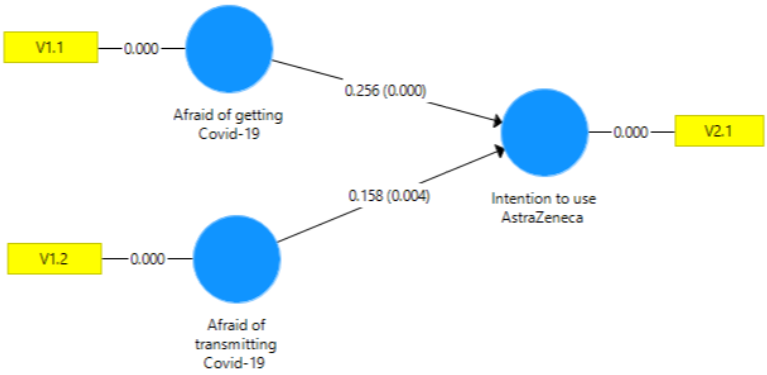


Figure 2. Fear of covid-19. CanSino (China).

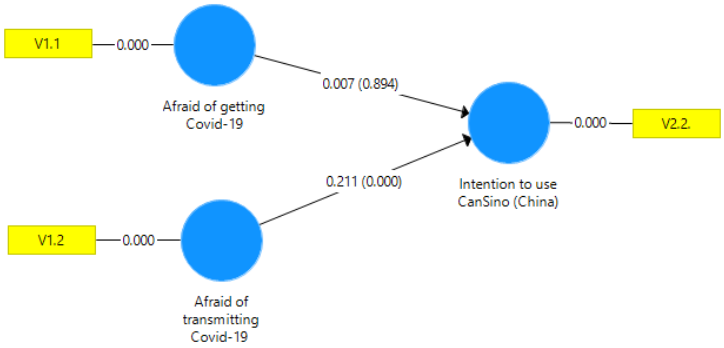


Figure 3. Fear of covid-19. Sputnik (Russia).

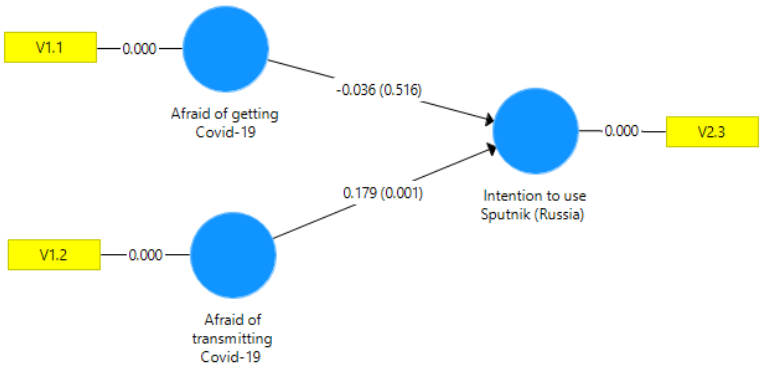


Figure 4. Fear of covid-19. Moderna (EEUU).

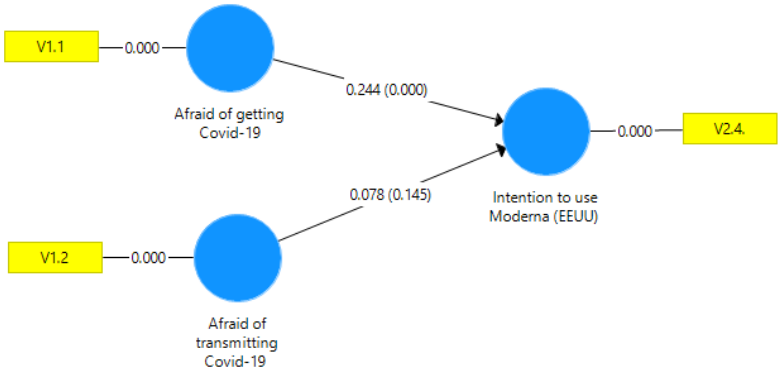


Figure 5. AstraZeneca (UK-EEUU).

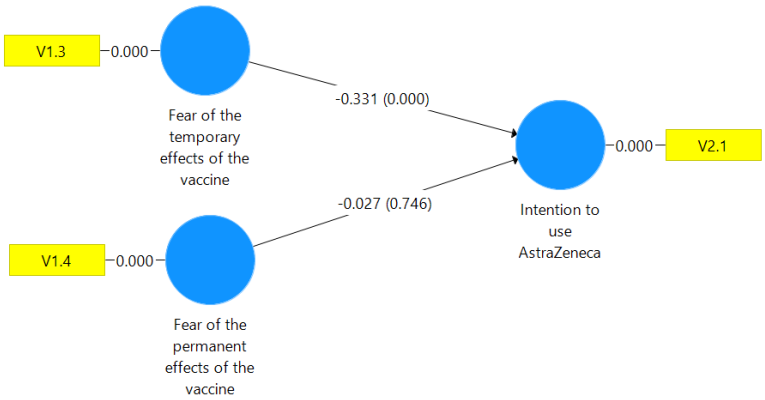


Figure 6. CanSino (China).

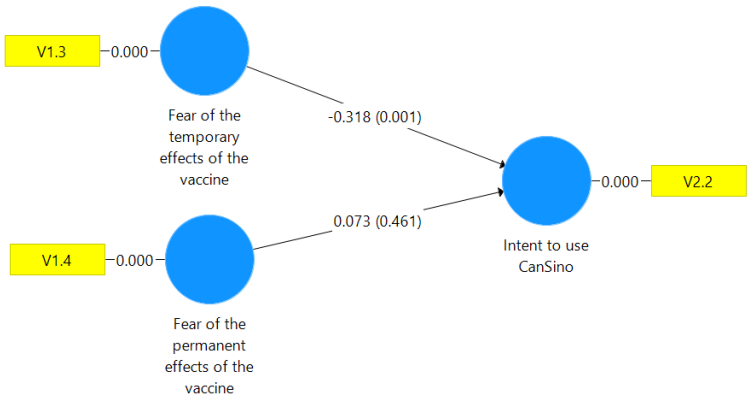


Figure 7. Sputnik (Russia).

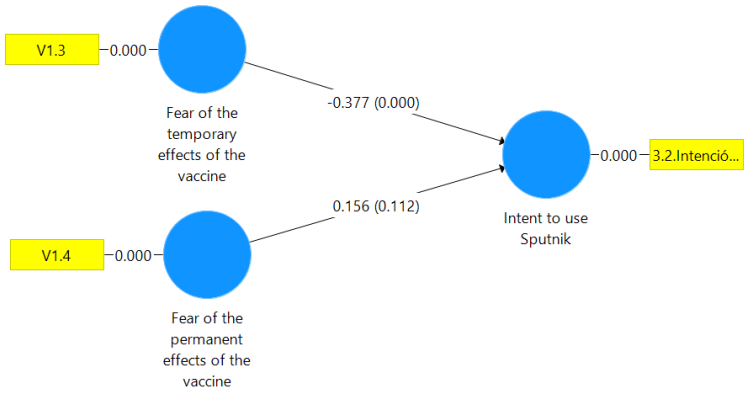
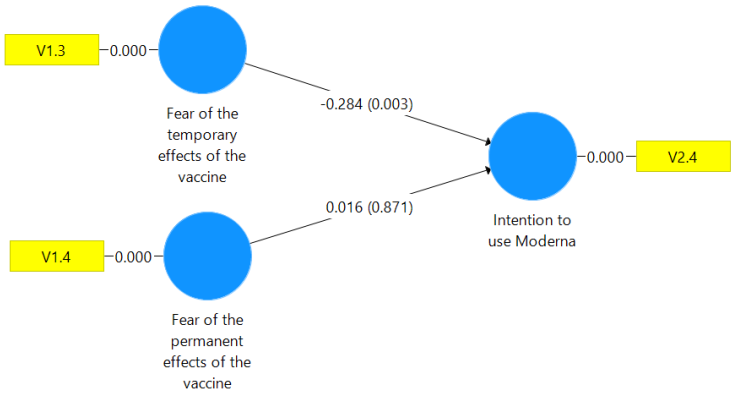


Figure 8. Moderna (EEUU).



Before analyzing the causal models, the reliability of each item has been verified (Hair et al., 2013), it has been confirmed that in all the models the standardized loadings were > 0.7 and the t-values were > 1.96.

In the present investigation it has not been necessary to analyze the reliability of the constructs since each observable variable has been treated independently. Nor has it therefore been necessary to analyze convergent validity or discriminant validity.

The results show that for the AstraZeneca vaccine (UK-USA) both fear of getting Covid-19 and fear of transmitting are predictors of the intention to use and also the ability of fear of Covid-19 to explain the intention to use the vaccines is the highest (R^2 = 0.138). Regarding the Moderna vaccine (USA), only fear of getting Covid-19 can explain the intention to use the vaccine. Besides, the ability of fear of Covid-19 to explain the intention to use these vaccines is the second highest (R^2 = 0.088). For the CanSino (China) and Sputnik (Russia) vaccines, the behavior is very similar: (i) the ability of fear of Covid-19 to explain the intention to use the vaccines is lower than for the other vaccines (R^2 Cansino = 0.046 and R^2 Sputnik = 0.026), and (ii) only fear of infecting others has an influence.

The results show that for the AstraZeneca vaccine (UK-USA) both fear of the temporary effect of the vaccine Covid-19 and fear of transmitting are predictors of the intention to use and also the ability of fear of Covid-19 to explain the intention to use the vaccines is the highest.

Regarding the effects of fear of the temporary and permanent effects of vaccines on the intention to get vaccinated in a similar way affect the intention to get vaccinated in all vaccines. While the fear of temporary effects significantly and negatively affects the intention to get vaccinated. The same does not happen with the permanent fear of effects, since in the latter case it does not significantly affect the intention to get vaccinated for all vaccines.

For the AstraZeneca Vaccine, the fear of the vaccine has an R^2 of 0.125. With small differences, the explanatory capacity of the models for the rest of the vaccines is similar: Moderna = 0.073, Cansino = 0.067, and Sputnik = 0.066.

CONCLUSIONS AND IMPLICATIONS

The study on the acceptance of vaccines for Covid-19 disease is considered essential to get people vaccinated. The question we can ask ourselves is: after receiving negative news about vaccines, does fear affect the acceptance of different vaccines in the same way?

To answer this question, after receiving the respondents' news published digitally about the possible serious side effects of a vaccine for Covid-19, we have verified the intention to be vaccinated with four different vaccines: AstraZeneca (UK-USA), CanSino (China), Sputnik (Russia) and Moderna (USA). We have also analyzed how fear of side effects affects the intention to get vaccinated with these vaccines and how fear of the Covid-19 disease affects the intention to get vaccinated with these vaccines.

The results have shown that we do not have the same perception about the different vaccines. In this sense, we have observed that the vaccines developed by countries with greater cultural ties with Spain (the United Kingdom and the United States) have greater acceptance than vaccines developed by countries with fewer cultural ties (China and Russia).

Our results have shown how side effects on vaccines affect respondents in the four vaccines analyzed. In all cases, the effects of fear of the temporary and permanent effects of vaccines on the intention to get vaccinated similarly affect the intention to get vaccinated. While the fear of temporary effects significantly and negatively affects the intention to get vaccinated. The same does not happen with the fear of the permanent effects that the vaccine can produce, since in the latter case, for all vaccines, it does not significantly affect the intention to get vaccinated. This is logical since while the temporary effects may not matter to some and matter more to others and therefore affect the explanation of the intention to use a vaccine, we may have a more similar opinion in the way in which we permanent (long-term) effects matter and therefore will not have explanatory power in the intention to be vaccinated.

We have also observed that the explanatory capacity of fear of the disease in the intention to get vaccinated is higher in vaccines designed in the United States and the United Kingdom than in those designed in China and Russia. However, the influence of the fear of contracting the disease and the fear of infecting others in the intention to get vaccinated has been different for the different vaccines. While for AstraZeneca (UK-USA) vaccine has influenced both the fear of suffering from the disease and the fear of transmitting it, for Moderna (USA) vaccine the influence is solely on having the disease. For CanSino (China) and Sputnik (Russia) vaccines, the influence only occurs for fear of transmitting the disease.

We can say that there is more similarity in terms of fear of the effects of vaccines than in terms of fear of the disease. Furthermore, as predicted by the previous literature, fear of the vaccine negatively affects the intention to be vaccinated and fear of the disease has positively affected the intention to be vaccinated (in cases where the influence is significant).

In conclusion, we have observed that as we have more interest in using a vaccine, fear becomes more important in the decision to vaccinate. That is, if a person has a low intention to use a vaccine, the fear caused by the news is of little importance to that person. But when you have a greater interest in using a vaccine, the fear caused by the news takes on greater importance.

The research has some limitations. On the one hand, the sample is only Spanish and therefore we do not know if it affects the same way for samples from other countries. It would be advisable to replicate the study for other countries. On the other hand, the news only referred to serious adverse side effects in a vaccine. Future research should analyze different types of news. In addition, it would be convenient to analyze the effect without news and with news for the same sample, which would allow us to observe the differences in the way the news affects the news.

ACKNOWLEDGEMENTS

This research was funded by the bridge grants for research projects awarded by the University of La Rioja (PROYECTOS PUENTE PP-2020-02), subsidized by Banco Santander, the COBEMADE research group at the University of La Rioja.

KEYWORDS: digital influence, ethical, fear to Covid-19, vaccine acceptance.

REFERENCES

Abebe, A. M., Mengistu, T., & Mekuria, A. D. (2019). Measles case, immunization coverage and its determinant factors among 12-23-month children, in Bassona Worena Woreda, Amhara Region, Ethiopia, 2018. *BMC Research Notes*, 12(1), 1-6. https://doi.org/10.1186/s13104-019-4104-8

Anraad, C., Lehmann, B. A., Visser, O., van Empelen, P., Paulussen, T. G. W., Ruiter, R. A. C., … van Keulen, H. M. (2020). Social-psychological determinants of maternal pertussis vaccination acceptance during pregnancy among women in the Netherlands. *Vaccine*, 38(40), 6254-6266. https://doi.org/10.1016/j.vaccine.2020.07.047

Anraad, C., Lehmann, B. A., Visser, O., van Empelen, P., Paulussen, T. G. W., Ruiter, R. A. C., … van Keulen, H. M. (2020). Social-psychological determinants of maternal pertussis vaccination acceptance during pregnancy among women in the Netherlands. *Vaccine*, 38(40), 6254-6266. https://doi.org/10.1016/j.vaccine.2020.07.047

Borena, W., Luckner-Hornischer, A., Katzgraber, F., & Holm-von Laer, D. (2016). Factors affecting HPV vaccine acceptance in west Austria: Do we need to revise the current immunization scheme? *Papillomavirus Research*, 2(June), 173-177. https://doi.org/10.1016/j.pvr.2016.10.001

CDC. (2014, May 1). Vaccine Testing and the Approval Process. Retrieved September 19, 2020, from https://www.cdc.gov/vaccines/basics/test-approve.html

Chan, E. Y., Cheng, C. K., Tam, G., Huang, Z., & Lee, P. (2015). Knowledge, attitudes, and practices of Hong Kong population towards human A/H7N9 influenza pandemic preparedness, China, 2014. *BMC public health*, 15(1), 1-10.

Chapman, G. B., & Coups, E. J. (2006). Emotions and preventive health behavior: worry, regret, and influenza vaccination. *Health psychology*, 25(1), 82.

Cordoba-Sanchez, V., Tovar-Aguirre, O. L., Franco, S., Arias Ortiz, N. E., Louie, K., Sanchez, G. I., & Garces-Palacio, I. C. (2019). Perception about barriers and facilitators of the school-based HPV vaccine program of Manizales, Colombia: A qualitative study in school-enrolled girls and their parents. *Preventive Medicine Reports*, 16(February), 100977. https://doi.org/10.1016/j.pmedr.2019.100977

Dijkstra, T. K., & Henseler, J. (2015). Consistent Partial Least Squares Path Modeling. *MIS Quarterly*, 39(2), 297-316. https://doi.org/10.25300/MISQ/2015/39.2.02

Dubé, E., Gagnon, D., Ouakki, M., Bettinger, J. A., Witteman, H. O., MacDonald, S., … Greyson, D. (2018). Measuring vaccine acceptance among Canadian parents: A survey of the Canadian Immunization Research Network. *Vaccine,* 36(4), 545-552. https://doi.org/10.1016/j.vaccine.2017.12.00

Fu, L. Y., Zimet, G. D., Latkin, C. A., & Joseph, J. G. (2017). Associations of trust and healthcare provider advice with HPV vaccine acceptance among African American parents. *Vaccine,* 35(5), 802-807. https://doi.org/10.1016/j.vaccine.2016.12.045

Gefen, D., Rigdon, E. E., & Straub, D. (2011). Editor's Comments: An Update and Extension to SEM Guidelines for Administrative and Social Science Research. *Mis Quarterly*, 35(2), iii-xiv. https://doi.org/10.2307/23044042

Han, S., Lerner, J., & Keltner, D. (2006). Feelings and consumer decision making. *Journal of Consumer Psychology*, 17(3), 158-168. https://doi.org/10.1126/science.1125877

Kyaw, W. M., Chow, A., Hein, A. A., Lee, L. T., Leo, Y. S., & Ho, H. J. (2019). Factors influencing seasonal influenza vaccination uptake among health care workers in an adult tertiary care hospital in Singapore: A cross-sectional survey. *American Journal of Infection Control*, 47(2), 133-138. https://doi.org/10.1016/j.ajic.2018.08.011

Lau, J. T. F., Yang, X., Tsui, H., & Kim, J. H. (2003). Monitoring community responses to the SARS epidemic in Hong Kong: from day 10 to day 62. *Journal of Epidemiology & Community Health*, 57(11), 864-870.

Lazarus, J. V., Ratzan, S., Palayew, A., Gostin, L. O., Larson, H. J., Rabin, K., ... & El-Mohandes, A. (2020). Hesitant or not? A global survey of potential acceptance of a COVID-19 vaccine*. medRxiv.* DOI: 10.1101/2020.08.23.20180307

Luz, P. M., Brown, H. E., & Struchiner, C. J. (2019). Disgust as an emotional driver of vaccine attitudes and uptake? A mediation analysis. *Epidemiology & Infection*, 147.

Malik, A. A., McFadden, S. M., Elharake, J., & Omer, S. B. (2020). Determinants of COVID-19 Vaccine Acceptance in the US*. medRxiv*. http://doi.org/10.1016/j.eclinm.2020.100495

Maltezou, H. C., Pelopidas Koutroumanis, P., Kritikopoulou, C., Theodoridou, K., Katerelos, P., Tsiaousi, I., … Loutradis, D. (2019). Knowledge about influenza and adherence to the recommendations for influenza vaccination of pregnant women after an educational intervention in Greece. *Human Vaccines and Immunotherapeutics*, 15(5), 1070-1074. https://doi.org/10.1080/21645515.2019.1568158

Manca, T. (2018). "One of the greatest medical success stories:" Physicians and nurses' small stories about vaccine knowledge and anxieties. *Social Science & Medicine*, 196, 182-189.

Nguyen, T. T. M., Lafond, K. E., Nguyen, T. X., Tran, P. D., Nguyen, H. M., Ha, V. T. C., … McFarland, J. W. (2020). Acceptability of seasonal influenza vaccines among health care workers in Vietnam in 2017. *Vaccine*, 38(8), 2045-2050. https://doi.org/10.1016/j.vaccine.2019.12.047

O'Grady, S. (2020, July 3). Why is coronavirus making the economy worse? - The Washington Post. Retrieved August 25, 2020, from https://www.washingtonpost.com/world/2020/07/03/how-has-coronavirus-pandemic-affected-global-poverty

Ortony, A., & Turner, T. J. (1990). What's Basic about Basic Emotions? *Psychological Review*, 97(3), 315-331. https://doi.org/10.1037//0033-295x.97.3.315

Otieno, N. A., Otiato, F., Nyawanda, B., Adero, M., Wairimu, W. N., Ouma, D., … Verani, J. R. (2020). Drivers and barriers of vaccine acceptance among pregnant women in Kenya. *Human Vaccines and Immunotherapeutics*, 00(00), 1-9. https://doi.org/10.1080/21645515.2020.1723364

Patil, S. S., Patil, S. R., Ganla, A., & Durgawale, P. M. (2020). Knowledge and awareness about cervical cancer and human papilloma virus (HPV) vaccine among nursing students. *Journal of Critical Reviews*, 7(12), 384-393. https://doi.org/10.31838/jcr.07.12.73

Pelegrín-Borondo, J. P., González-Menorca, C. G., & Meraz, L. (2019). The influence of the emotions produced by the wine offer, winery visits, and wine news on wine purchase intent in tourists. *Spanish journal of agricultural research*, 17(1), 4.

Pelegrín-Borondo, J., Arias-Oliva, M., & Olarte-Pascual, C. (2017). Emotions, price and quality expectations in hotel services. *Journal of Vacation Marketing*, 23(4), 322-338. https://doi.org/10.1177/1356766716651305

Pelegrín-Borondo, J., Juaneda-Ayensa, E., González-Menorca, L., & González-Menorca, C. (2015). Dimensions and basic emotions: A complementary approach to the emotions produced to tourists by the hotel. *Journal of Vacation Marketing*, 21(4), 351-365. https://doi.org/10.1177/1356766715580869

Poland, G. A. (2010). The 2009-2010 influenza pandemic: effects on pandemic and seasonal vaccine uptake and lessons learned for seasonal vaccination campaigns. *Vaccine*, 28, D3-D13.

Quinn, S. C., Jamison, A. M., An, J., Hancock, G. R., & Freimuth, V. S. (2019). Measuring vaccine hesitancy, confidence, trust and flu vaccine uptake: Results of a national survey of White and African American adults. *Vaccine*, 37(9), 1168-1173. https://doi.org/10.1016/j.vaccine.2019.01.033

Ram, J., Corkindale, D., & Wu, M.-L. (2014). ERP Adoption and the Value Creation: Examining the Contributions of Antecedents. *Journal of Engineering and Technology Management*, 33, 113-133. https://doi.org/10.1016/j.jengtecman.2014.04.001

Russell, J. A. (2003). Core Affect and the Psychological Construction of Emotion. Psychological Review, 110(1), 145-172. https://doi.org/10.1037/0033-295X.110.1.145

Russell, J. A. (2009). Emotion, Core Affect, and Psychological Construction. *Cognition and Emotion*, 23(7), 1259-1283. https://doi.org/10.1080/02699930902809375

Russell, J. A., & Barrett, L. F. (1999). Core Affect, Prototypical Emotional Episodes, and Other Things Called Emotion: Dissecting the Elephant. *Journal of Personality and Social Psychology*, 76(5). https://doi.org/10.1037//0022-3514.76.5.805

Sarathchandra, D., Navin, M. C., Largent, M. A., & McCright, A. M. (2018). A Survey Instrument for Measuring Vaccine Acceptance. *Preventive Medicine*, 109, 1-7. https://doi.org/10.1016/j.ypmed.2018.01.006

Scherer, K. R. (2005). What are Emotions? And how can they be Measured? *Social Science Information*, 44(4), 695-729. https://doi.org/10.1177/0539018405058216

WHO. (2020, August 25). Coronavirus disease (COVID-19). Retrieved August 25, 2020, from https://www.who.int/emergencies/diseases/novel-coronavirus-2019

ASSESSMENT OF ETHICAL ON THE INTENTION TO USE OF WEARABLES

Eva Reinares-Lara, Isabel Álvarez, Mario Arias-Oliva, Cristina Olarte-Pascual, Maria Alesanco-Llorente

Universidad Rey Juan Carlos (Spain), Universidade Autónoma de Lisboa (Portugal), Universidad Complutense de Madrid (Spain), Universidad de La Rioja (Spain), Universidad de La Rioja (Spain)

eva.reinares@urjc.es, ialvarez@autonoma.pt; mario.arias@ucm.es;cristina.olarte@unirioja.es; maria.alesanco@alum.unirioja.es

INTRODUCTION

The human body has become a support for intelligent technologies as wearables (externally-worn intelligent devices, such as watches, bracelets, clothing, glasses and headphones) that has opened an ethical debate about their development, commercialisation, and use in modern society. Wearables represent an expanding market and forecasts for the use indicate that the world market will reach 279 million units in 2023, with a compound annual growth rate of 8.9%, being the health field a large market (IDC, 2019).

In the literature, this scenario poses ethical dilemmas about wearable in a variety of areas, including the social, economic, environmental, educational, moral and philosophical (e.g. Shipp et al., 2014; Li, 2015; Mok et al., 2015; Ferenbok et al., 2016; Hofmann et al., 2017; Segura Anaya et al., 2017; McCall et al., 2019; Kostick et al., 2019; Kreitmair, 2019; Olarte-Pascual, 2021). Although the ethical issues have been approached from institutional and organisational perspectives, fundamentally using discourse methodology, few studies have taken a demand perspective about the influence of ethics on the acceptance and intention of using wearables (e.g. Hofmann et al., 2017; Segura Anaya et al., 2017).

The present study addresses this research gap by modelling the acceptance of capacity-enhancing wearable ITDs, using "ethical judgment" as an antecedent of intention to use. Ethical judgment has been defined as a cognitive process in which the individual must "judge which course of action is morally right" (Nguyen & Biderman, 2008, p. 628).

The results of this work advance the theoretical development of ethics as applied to the acceptance of new technologies: ethical judgment is key for the acceptance of wearables? At the same time, the demand approach will establish operational implications that can help, while taking account of users' ethical judgments, guide the development and commercialisation of capacity-enhancing wearables.

THE INFLUENCE OF ETHICS ON THE ACCEPTANCE OF WEARABLE

As previously noted, a fundamental criterion for the acceptance of wearable is the ethical assessment of these technologies. In discussing, Ferenbok et al. (2016, p. 95) stated that *"wearable devices represent more than just a potential economic disruption, but, in a broader sense, a disruption of the ethics by which we live"*. Disruptive technologies, through a process of refinement, improvement and innovation, create new standards (Christensen et al., 2018). Ethics allow the controversy between the potential benefits that can be achieved through technological progress, and the duty not to endanger this progress, to be addressed.

In the framework of ethical judgment, ethical evaluations of actions have been conceptualised as individual cognitive processes (Nguyen & Biderman, 2008). In turn, *psychological contract theory* conceptualises decision-making subjectively (Thompson & Hart, 2006). This theoretical basis can be used to address similar decisions made by individuals in the absence of absolute rules of what one can and cannot do (Goel et al., 2016). Decisions and actions are often guided by applied ethical perceptions, rather than a complete understanding of what can or should be done (LaFollette, 2002; Cohen & Wellman, 2005). In the sphere of circular evolutionary ethics, what an individual considers ethical influences his/her behaviour and, over time, the behaviours they observe influence what they believe to be ethical (Goel et al., 2016). In the present study, we believe it is appropriate to analyse the impact of ethics on intention to use wearables on the basis of individuals' perceptions of what behaviours, from an applied ethical viewpoint, are appropriate (Thompson & Hart, 2006).

Reidenbach and Robin (1990) argued that individuals use more than one reason to make ethical judgments, and thus they established the multidimensional ethics scale (MES) used in the literature to explain the influence of ethical judgment on people's behavior. Shawver and Sennetti (2009) proposed a new scale, which they called the Composite MES; this has five dimensions, "moral equity", "relativism", "utilitarianism", "egoism" and "contractualism" (deontology). The Composite MES has been widely used to explain the impact of ethical judgments on behaviour (e.g. Mudrack & Mason, 2013; Manly et al., 2015; Kara et al., 2016). To a lesser extent, the MES has been used in the context of consumption behaviour (e.g. Nguyen & Biderman, 2008; Jones & Leonard, 2016; Leonard & Jones, 2017); however, in the field of wereables acceptance, the influence of ethical judgments and the Composite MES dimensions have been discussed only by Olarte-Pascual (2021):

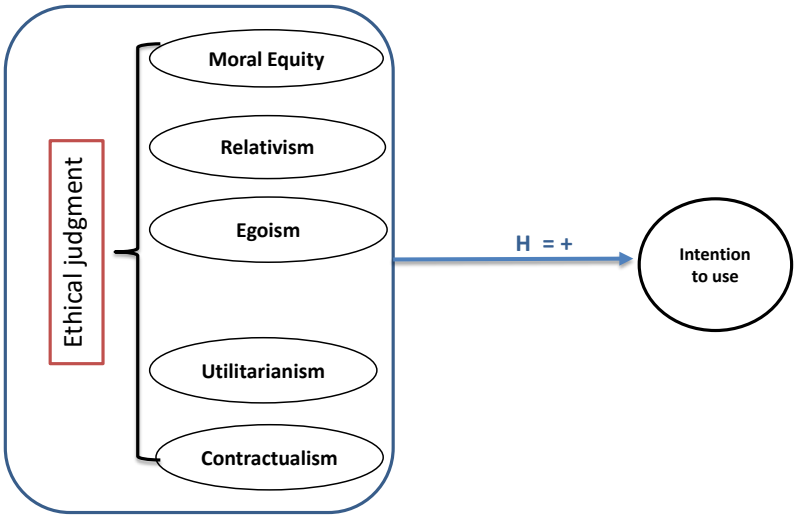
- The "moral equity" dimension refers to the "*individual perception of fairness and justice as well as what is right and wrong in its broadest sense*" (Nguyen & Biderman, 2008, p. 628). According to Leonard et al. (2017), this dimension encompasses fairness, justice, rightness, and goodness. Hofmann et al. (2017) concluded that, in relation to smart glasses, justice appears to be crucial for the successful development, evaluation, decision-making, implementation, use, and formation of knowledge and norms**.** Weber and Zink (2014) found that the use of smart glasses and other intelligent devices increases the digital divide and, in the sports field, Bozyer (2015) showed that these devices create unfair advantages for those who have access to the technologies because they can, in consequence, train more effectively. Wearables can have negative consequences, such as the creation of a social divide between those who can afford the latest innovative technology and those who cannot.
- "Relativism" refers to the perception that what is correct is based on guidelines/parameters embedded in social/cultural systems, rather than on individual considerations (Reidenbach & Robin, 1990; Nguyen & Biderman, 2008). Ferenbok et al. (2016) found that the most modern wearable computers, such as smart glasses, offer unprecedented portability and ability to capture images, and thus can go where no digital eye has gone before, which represents a departure from established social norms.
- "Utilitarianism" has been defined as "*an action based on cost and benefit analyses, such that the action will bring about the greatest good for the greatest number*" (Nguyen & Biderman, 2008, p. 628). It must be kept in mind that capacity-enhancing technologies might help in societal advancement and that impeding them could be considered unethical (Berger et al., 2008). Wearables improve the quality of life of their users, promote lifestyle changes and save time and money (Segura Anaya et al., 2017). These authors also found that, while wearables have considerable benefits, device dependency, and privacy and security concerns, are major

challenges. In light of these potential problems, the balance between the costs and the societal benefits of technology must be analysed from a utilitarian perspective.

- "Egoism'" has been defined as acting in a manner that promotes only one's own long-term self-interest (Nguyen & Biderman, 2008). This dimension focuses on the consequences for the individual (Reidenbach & Robin, 1990), and Leonard et al. (2017) concluded that the individual's intention to behave ethically is driven by the benefits that the behaviour will bring to him/her. Wearable technologies have been developed to improve, increase, and empower individuals. For example, smart glasses empower and improve their users' cognitive capacities, although this may provoke negative reactions in others (Hofmann et al., 2017).
- "Contractualism" (deontology) refers to the "*individual perception of what is right versus wrong based on notions of an implied contract that exists between business and society*" (Nguyen & Biderman, 2008, p. 633). Reidenbach and Robin (1990) argued that this dimension reflects the deontological concept and encompasses notions of implicit obligations, contracts, duties and rules. Shipp et al. (2014) and Mok et al. (2015) examined the role of wearable cameras in the research field and evaluated the amount of data they provided to researchers and the related ethical concerns in regard to respect for personal autonomy, common well-being, trust in third parties, anonymity, confidentiality, privacy, beneficence (responsibility to do good) and non-maleficence (responsibility to avoid doing harm). According to Thierer (2015), societal and individual adaptation play key roles in the acceptance of wearables. Although great privacy and security challenges await, individuals and institutions will adjust in an evolutionary, resilient fashion, just as they have to earlier disruptive technologies.

Based on this theoretical background, the authors propose to advance the knowledge of the impact of ethical judgment and its dimensions on digital natives' intention to use wearable. Few researchers have discussed its influence on the acceptance of wearable. Thus, a working hypothesis is proposed: *H. Ethical judgment (moral equity, relativism, egoism, utilitarianism, contractualism) positively affects intention to use wearable.*

Figure 1. Structural model.



METHOD

To test the proposed hypotheses an online survey was undertaken with an international sample of 1,563 digital-native higher education students who assessed levels of technological competence from America, Asia and Europe. The characteristics of the sample are: Men 47.1% and Women 52.9%; Average 21.5 years and 19.39% wearing a wearable (e.g. bracelet, watch, glasses, or intelligent clothes).

The scale used to measure Ethical judgement have been adapted from Composite MES by Shawver & Sennetti (2009) with a semantic differential - 5 to +5 and intention to use with a 11-point Likert scale adapted from Venkatesh & Davis (2000).

The collected data were analysed through structural equation modelling (SEM), specifically, using the Consistent Partial Least Square (PLSc) technique.

To test the proposed hypotheses a sequential 3-step statistical process was followed: 1) Assessment of the measurement model. The measurement model was assessed by verifying the reliability and validity of the measurement scales; 2) Assessment of the structural model. For the model the R^2, path coefficients, and their significance were estimated.

RESULTS

Assessment of the measurement model

The results of the PLSc SEM analyses indicated that in the model, the standardised loadings of the observable variable "self-promoting for me" were lower than 0.7. This variable belongs to the egoism factor. Furthermore, in the analysis of the discriminant validity of the egoism and utilitarianism factors, it was observed that the heterotrait-monotrait ratio (HTMT) criterion was higher than 0.9. This led us to eliminate the observable variable “self-promoting for me” from the model. The results, without this variable, are presented below. All standardised loadings are higher than .7 and all t-values are higher than 1.96 (Hair et al., 2013), so the reliability of the indicator is verified (see Table 1). To check for common method bias the partialling out “marker” variable method recommended by Podsakoff et al. (2003) was used, following the process suggested by Tehseen et al. (2017). A marker variable was introduced as a predictor for the endogenous constructs of the wearable model. Subsequently, the R^2 values of the endogenous constructs before and after adding the marker variable were examined. The results showed similar R^2 values before and after introducing the marker variable. This result establishes that there is no substantial common method bias.

Table 1. Standardised loading values (t-values) of the dimensions of ethical judgement and intention to use wearables.

Moral Equity (ME)	
Unjust/Just	0.85 (33.1)
Unfair/Fair	0.87 (32.9)
Not morally right/Morally right	0.85 (35.3)
Relativism (REL)	
Not acceptable to my family/Acceptable to my family	0.89 (29.4)
Culturally unacceptable/Culturally acceptable	0.75 (22.2)
Traditionally unacceptable/Traditionally acceptable	0.79 (22.0)
Egoism (EGO)	
Not self-promoting for me/Self-promoting for me	single item

Utilitarianism (UTI)	
Produces the least utility/Produces the greatest utility	0.85 (36.5)
Minimise benefits and maximise hurt/ Maximise benefits and minimise hurt	0.76 (34.2)
Contractualism (CON)	
Violates/does not violate an unwritten contract	0.96 (51.8)
Violates/does not violate an unspoken promise	0.91 (45.5)
Intention to use wearables (IU)	
Intention to use	0.92 (76.9)
Use prediction	0.91 (71.3)

Table 2 shows that the reliability was adequate: Cronbach's *alpha* and composite reliability returned values greater than 0.7. Convergent validity is also verified as the average variance extracted (AVE) was greater than 0.5. Similarly, the discriminant validity criterion was met: the HTMT values were correct in all cases and the square root of the AVEs was greater than the inter-construct correlations (Roldán & Sánchez-Franco, 2012).

Table 2. Composite reliability, Cronbach's *alpha*, AVE (convergent validity) and discriminant validity.

Construct	**Composite reliability > 0.7**	**Cronbach's *alpha***	**AVE > 0.5**	**ME**	**REL**	**EGO**	**UTI**	**CON**	**IU**
ME	0.89	0.89	0.73	**0.86**	0.86	0.67	0.79	0.73	0.49
REL	0.85	0.85	0.66	0.86	**0.81**	0.62	0.79	0.77	0.45
EGO	1.00	1.00	1.00	0.67	0.62	**1.00**	0.80	0.60	0.57
UTI	0.79	0.79	0.65	0.79	0.78	0.79	**0.81**	0.78	0.64
CON	0.93	0.93	0.87	0.73	0.76	0.60	0.77	**0.93**	0.41
IU	0.90	0.90	0.83	0.49	0.45	0.57	0.64	0.41	**0.91**

Note: The diagonal elements (in bold) are the square root of the AVEs. The off-diagonal elements are the inter-construct correlations. The elements above the diagonal (in bold) are the HTMT values.

Assessment of the structural model

Table 3 shows the R^2 and Q^2 values, the path coefficients (direct effects), the path coefficients (direct effect), the Student´s t-test values and p-values for each antecedent variable of intention to use wearables, from which the influence of ethical judgment on intention to use can be extracted.

The R^2 for the intention to use wearables model is 0.44. The Q^2 provided by PLS Predict was greater than 0. This indicates that the exogenous variables, indeed, predict the endogenous variable. The results confirm that the model have predictive relevance.

Table 3. Effect on the endogenous variables.

	R^2	**Q^2**	**Path coefficient**	**Student's *t*-test**	***p*-value**
IUW	44.4%	0.18			
ME=> (+) IUW			0.10	1332	0.18
REL=>(+) IUW			-0.13	1652	0.10
EGO=>(+) IUW			0.15	2240	0.03
UTI=>(+) IUW			0.70	5960	0.00
CON=>(+) IUW			-0.19	3437	0.00

Note: Based on one-tailed Student's t-distribution (4.99).

The ethical judgment dimensions “egoism” and “utilitarianism” significantly influenced intention to use warables devices. "Moral equity" and “relativism” did not influence intention to use wearables. “Contractualism” affected significantly but negatively.

CONCLUSIONS

In the framework of new technology acceptance, the principal conclusion to be drawn from this work is that the ethical judgment construct has high explanatory power for digital natives’ intention to use new capacity-enhancing wearable technologies (R^2 = 44.4%; Q^2 = 0.184).

“Utilitarianism” is the most important dimension for wearables. We argue that that when more is known about new devices, such as wearables, ethical judgments focus more on whether they are useful to society in terms of their benefits (improved quality of life, lifestyle changes, time and money savings) vs their associated costs and inconvenience (device dependency, privacy and security concerns, among others) (Segura Anaya et al., 2017), leaving other ethical aspects in the background, as the devices have already been assimilated and the objections overcome in the framework of circular evolutionary ethics.

The effect of “relativism” is negative, although not significant (-13.3%) for intention to use wearables. "Relativism" is based on the idea that "social and cultural systems are important in helping us define our ethical beliefs" (Reidenbach & Robin, 1990, p. 646). In this sense, for a known product, which has been socially accepted, and is not perceived as bodily invasive, such as wearables, this spread of opinion is not observed in digital natives (Pelegrín-Borondo et al., 2017).

The "moral equity" ethical judgment dimension had a positive influence on intention to use of wearables. However, was not significant probably because digital natives’ ethical judgment is almost all explained by the “utilitarianism” dimension. This result contrasts with those of Hofmann et al. (2017) and Weber and Zink (2014), which showed that smart glasses and other intelligent devices widen the digital divide, which raises the issue about whether such a gap is morally fair.

The influence of “contractualism” on intention to use wearables is negative and significant. In other words, contrary to expectations, "contractualism", the implicit contract that exists between the individual and society, inversely influences intention to use wereables. We can say that, although wearables are a known quantity, and that their use is expected to increase significantly in the coming years (Hayward, 2018; IDC, 2019), there is no defined social norm in favour of, or against, these devices, so the related social pressure might be positive or negative.

The results of this study allow us to establish a series of operational implications to guide and design the responsible development and commercialisation of wearables.

Ethical aspects are key in explaining digital natives’ acceptance of wearables. The first implication is that any organisation that wishes to participate and compete in this sector must develop an ethical strategy based on the ethical judgments of these users.

It is also important for organisations to know how intention to use can be strengthened by addressing the dimensions of ethical judgment. The intensity and the direction of the effects of the five ethical judgment dimensions on intention to use wearables suggest the following practical implications will help promote their acceptance, and prevent their rejection, by digital natives:

Companies should focus their efforts on the “utilitarianism” dimension, which explains most of intention to use wearables, due to its high explanatory power. To promote the use of wearables, the marketing community must continue to emphasise their utility for society, in line, for example, with

the benefits reported by Segura Anaya et al. (2017), in terms of improved quality of life, and the optimisation of productivity and economic and time resources. On the other hand, to prevent rejection, the marketing axis must focus on reducing or eliminating societal perceptions of privacy and security problems related to the use of these devices. In addition, due to the importance of the "utilitarianism" dimension, public powers must guarantee the rights of individuals in this matter. Kreitmair (2019) evaluated the ethical dimensions of wearable technologies based on their contribution to the "good life" of the user, in accordance with the Aristotelian "human flourishing" concept, and argued that, as a criterion of consumption ethics, that the momentum of utilitarianism should not be arrested. However, McCall et al. (2019) noted the widespread marketing claims promoting the use of wearables, promising health, personal cognitive and well-being benefits, absent of any warnings about possible risks and side effects. We believe it is necessary that, in line with the ethical considerations of Kostick et al. (2019) about the effects of these type of claims, to guarantee informed choice the commercial exploitation of the "utility" dimension be supported by scientific evidence.

The study of the key variables of the acceptance of wearables is an insufficiently studied field, which is related to a high growth potential and a high impact on job creation and economic activity. A future research line could identify new variables that should expand the models of acceptance of wearables and analyse whether the specific applications of wearables and the contexts of use moderate the explanatory power of the ethical judgment model and how.

KEYWORDS: Wearables, Technology acceptance, Ethical judgment, Composite MES.

REFERENCES

Bozyer, Z. (2015). Augmented Reality in Sports: Today and Tomorrow. International Journal of Science Culture and Sport, 3(4), 314-325.

Christensen, C. M., McDonald, R., Altman, E. J., & Palmer, J. E. (2018). Disruptive innovation: An intellectual history and directions for future research. *Journal of Management Studies*, 55(7), 1043-1078.

Cohen, A. I., & Wellman, C. H. (Eds.). (2005). *Contemporary Debates in Applied Ethics*. Wiley-Blackwell.

Ferenbok, J., Mann, S., & Michael, K. (2016). The Changing Ethics of Mediated Looking: Wearables, veillances, and power. *IEEE Consumer Electronics Magazine*, 5(2), 94-102.

Goel, L., Hart, D., Junglas, I., & Ives, B. (2016). Acceptable IS Use: Conceptualization and measurement. *Computers in Human Behavior*, 55, 322-328.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2013). Partial least squares structural equation modeling: Rigorous applications, better result and higher acceptance. *Long Range Planning*, 46(1/2), 1-12.

Hayward, J. (2018). Wearable Sensors 2018–2028: Technologies, Markets & Players. IDTechEx: Cambridge, UK. Retrieved from https://www.idtechex.com/de/research-report/wearable-sensors-2018-2028-technologies-markets-and-players/555

Hofmann, B., Haustein, D., & Landeweerd, L. (2017). Intelligent-glasses: exposing and elucidating the ethical issues. *Science and Engineering Ethics*, 23(3), 701-721.

IDC (2019). World Quarterly Wearable Device Tracker. International Data Corporation. International Data Corporation (IDC). Retrieved from https://www.idc.com/getdoc.jsp?containerId=prUS44930019

Jones, K., & Leonard, L. N. (2016). Applying the Multidimensional Ethics Scale in C2C E-Commerce. *Issues in Information Systems*, 17(1). pp. 26-36.

Kara, A., Rojas-Méndez, J. I., & Turan, M. (2016). Ethical Evaluations of Business Students in an Emerging Market: Effects of Ethical Sensitivity, Cultural Values, Personality, and Religiosity. *Journal of Academic Ethics*, 14(4), 297-325.

Kostick, K. M., Sierra-Mercado, D. & Lázaro-Muñoz, G. (2019). Ethical and Social Considerations for Increasing Use of DTC Neurotechnologies. *AJOB Neuroscience*, 10 (4), 183-185.

Kreitmair, K. V. (2019). Dimensions of ethical direct-to-consumer neurotechnologies. *AJOB Neuroscience*, 10(4), 152-166.

LaFollette, H. (2002). *Ethics in practice* (2nd ed.). Oxford: Blackwell Publishing.

Leonard, L. N., & Jones, K. (2017). Ethical Awareness of Seller's Behavior in Consumer-to-Consumer Electronic Commerce: Applying the Multidimensional Ethics Scale. *Journal of Internet Commerce*, 16(2), 202-218.

Leonard, L. N., Riemenschneider, C. K., & Manly, T. S. (2017). Ethical Behavioral Intention in an Academic Setting: Models and Predictors. *Journal of Academic Ethics*, 15(2), 141-166.

Li, H. (2015). Emotional Design in Wearable Technology. Viitattu, 25, 1-5. Retrieved from https://digitalwellbeing.org/wp-content/uploads/2015/11/2015-Emotional-Design-in-Wearable-Technology.pdf

Manly, T. S., Leonard, L. N., & Riemenschneider, C. K. (2015). Academic integrity in the information age: Virtues of respect and responsibility. *Journal of Business Ethics*, 127(3), 579-590.

McCall, I. C., Lau, C., Minielly, N., & Illes, J. (2019). Owning ethical innovation: Claims about commercial wearable brain technologies. *Neuron*, *102*(4), 728-731.

Mok, T. M., Cornish, F., & Tarr, J. (2015). Too much information: visual research ethics in the age of wearable cameras. *Integrative Psychological and Behavioral Science*, 49(2), 309-322.

Mudrack, P. E., & Mason, E. S. (2013). Ethical judgments: what do we know, where do we go? *Journal of Business Ethics*, 115(3), 575-597.

Nguyen, N. T., & Biderman, M. D. (2008). Studying ethical judgments and behavioral intentions using structural equations: Evidence from the multidimensional ethics scale. *Journal of Business Ethics*, 83(4), 627-640.

Olarte Pascual, C., Pelegrín Borondo, J., Reinares Lara, E. & Arias Oliva, M. (2021). From wearable to insideable: Is ethical judgment key to the acceptance of human capacity-enhancing intelligent technologies? *Computers in Human Behavior*. 114 (january), 1-11.

Pelegrín-Borondo, J., Arias-Oliva, M., Murata, K., & Souto-Romero, M. (2020). Does Ethical Judgment Determine the Decision to Become a Cyborg? *Journal of Business Ethics*, 161(1), 5-17.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psycholog*y, 88(5), 879-903.

Reidenbach, R. E., & Robin, D. P. (1990). Toward the development of a multidimensional scale for improving evaluations of business ethics. *Journal of Business Ethics*, 9(8), 639–653.

Roldán, J.L., & Sánchez-Franco, M.J. (2012). Variance-based structural equation modeling: Guidelines for using partial least squares in information systems research. In M. Mora, O. Gelman, A.

Steenkamp, & M. Raisinghani (Eds.), *Research methodologies, innovations and philosophies in software systems engineering and information systems* (pp.193–222). Hershey, PA: Raisinghan Information Science Reference.

Segura Anaya, L. H., Alsadoon, A., Costadopoulos, N., & Prasad, P. W. C. (2017). Ethical Implications of User Perceptions of Wearable Devices. *Science and Engineering Ethics*, 24(1), 1–28.

Shawver, T.J., & Sennetti, J.T. (2009). Measuring Ethical Sensitivity and Evaluation. *Journal of Business Ethics*, 88(4), 663–678.

Shipp, V., Skatova, A., Blum, J., & Brown, M. (2014, May). The ethics of wearable cameras in the wild. In *Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology* (p. 18). Chicago, USA: IEEE Press.

Tehseen, S., Ramayah, T., & Sajilan, S. (2017). Testing and controlling for common method variance: A review of available methods. *Journal of Management Sciences*, 4(2), 142-168.

Thierer, A. D. (2015). The internet of things and wearable technology: addressing privacy and security concerns without derailing innovation. *Richmond Journal of Law and Technology*, 21(2), 1-118.

Thompson, J., & Hart, D. (2006). Psychological contracts: a nano-level perspective on social contract theory. *Journal of Business Ethics*, 68(3), 229-241.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science*, 46, 186-204.

Weber, H., & Zink, K. J. (2014). Boon and Bane of ICT Acceleration for Vulnerable Populations. In C. Korunka, & P. Hoonakker (Eds.), *The Impact of ICT on Quality of Working Life* (pp. 177-190). Dordrecht: Springer.

Weston, M. (2015). Wearable surveillance – a step too far? *Strategic HR Review*, 14(6), 214-219.

5. Marketing, Technology and Ethics

WOKE WASHING IN THE WAKE OF COVID-19: A CASE STUDY ON AMAZON

Alice Gina Mary Crowley

University of South Australia (Australia)

alice.crowley@mymail.unisa.edu.au

ABSTRACT

Although many businesses have seen detrimental economic impacts as a result of Covid-19, Amazon is one retailer that has thrived in the face of pandemic. CEO Jeff Bezos increased his personal wealth by US$70bn since the start of the pandemic, and in the third quarter of 2020, the company earned US$6.3bn – the highest quarterly earnings in Amazon's 26 year history. However, Amazon has come increasingly under fire for allegations of employee mistreatment, including claims that they underpay their staff and making them work in dangerous and unsafe conditions – which has been exasberated as a result of the pandemic. This paper conducts a qualitative content analysis on 40 tweets that Amazon posted between 1st April to 31 December 2020 to examine how the retailer framed and referred to their employees, and potentially engaged in woke washing by presenting themselves as being concerned with their employees' wellbeing when their practices say differently. In addition, to provide insight on how users percieve these frames and Amazon's treatment of employees more broadly, this study also conducts a qualitive content analysis on 2843 tweet replies to these posts. This study finds that Amazon utilised frames which maintain oppressive employee-employer relationships, normalise the dangerous and unsafe working conditions that Amazon's employees face, and minimise the experiences of employees. Additionally, a significant amount of Twitter users called out Amazon for actively engaging in woke washing by creating an inaccurate representation of their employee relations on Twitter. This paper discusses the scope for future efforts, and labour campaigns specifically, to utilise Twitter to create better working conditions for Amazon staff.

INTRODUCTION

The magnitude of the global recession triggered by COVID-19 is unprecedented in modern times. In May 2020, the Asian Development Bank (2020) announced that the pandemic could cost the global economy between US$5.8 and US$8.8 trillion. Between April and June 2020, the United States' economy contracted 32.9%, the United Kingdom's contracted 20.4%, and Japan's contracted 27.8% (BBC News, 2020a; Dennis, 2020). Social and travel restrictions have also reduced the workforce in most economic sectors (Nicola et al., 2020). The hardest hit of these include hospitality, tourism, entertainment & arts, and sports & recreation (Wilkins, 2020). Social restrictions have also triggered the collapse of several well-known businesses, such as the Arcadia fashion group (owner of Topshop and Miss Selfridge), airline Virgin Australia and Swedish stationary giant Kikki.K (Davey, 2020; Nine News Australia, 2020; Virgin Australia, 2020).

However, one company which has reaped benefits from COVID-19 is online Seattle-based retailer Amazon. With lockdown measures causing people to stay home, people worldwide turned to e-commerce platforms as a means of purchasing products and services. In April 2020, *The Guardian* reported that Amazon customers were spending almost US$11,000 per second on the website. The company's founder and chief executive, Jeff Bezos, increased his personal wealth by US$70bn since the start of the pandemic, to US$185bn as of December 2020 (Neate, 2020). It has been widely

discussed that Bezos could give US$105,000 to every Amazon employee and would still be as wealthy as he was before the pandemic (Reich, 2020). In the third quarter of 2020, the company earned US$6.3bn – the highest quarterly earnings in Amazon's 26 year history (Rana & Dastin, 2020). It is therefore unsurprising that the retailer has been described as "Wall Street's biggest winner from coronavirus" (Randewich, 2020).

However, despite its financial success, Amazon has come under fire for its treatment of employees throughout the pandemic. The coronavirus outbreak has highlighted how much the economy depends on frontline workers (Crane & Matten, 2020). These frontline workers have been continually exposed to the virus, with some organisations failing to take the safety measures required to protect employees (The Lancet, 2020). This is especially true when it comes to Amazon, with the retailer being accused of lacking adequate deep cleaning after there were confirmed COVID-19 cases at its fulfilment centres (Sainato, 2020); failing to provide personal protective equipment, hand sanitiser and socially distanced workspaces (Pound, 2020); and a lack of paid sick leave (Woodward, 2020). Some Amazon employees shared their stories of mistreatment with the news media, arguing the company priorities gruelling productivity targets over employee safety.

Amazon has continually denied these allegations. In March 2020, when a New York worker organised a walkout to protest for what he argued were inadequate safety measures, he was fired by Amazon. A leaked memo illustrates that Amazon's general council wanted to portray the worker as "not smart or articulate" (Soper & Day, 2020; Wong, 2020). In response to growing accusations, the retailer launched television advertisements and a docuseries which portrayed Amazon as being deeply concerned with the safety of its 'retail heroes'. Amazon's Twitter page shows a number of posts which represent the retailer's employees as happy and well-supported, as well as highlighting the Covid-safe practices enforced within the company's facilities. When contrasting these social media posts to the allegations made by employees, it suggests that Amazon's internal behaviours may not align with the image it wishes to project. This is known as 'woke washing' – a form of inauthentic brand activism whereby a brand's practices do not align with their messaging.

The goal of this paper is to investigate how Amazon uses the values of employee health and safety to woke wash their online marketing campaign and to investigate users' responses to it. This study collects 40 posts tweeted by Amazon on Twitter between 1st April to 30th December 2020 which refers to their employees specifically. In addition, to understand how audiences perceived Amazon's posts, this paper collects 2833 tweet replies. A qualitative content analysis is conducted on Amazon's Twitter posts to investigate the frames the company used in representing itself and its employees whilst allegations of employee mistreatment accrued in the media. A qualitative content analysis is also conducted on users' Twitter replies to understand the themes online audiences used when responding to the campaign. The objective of this paper is therefore to offer a unique and topical case study on how Amazon has engaged in woke washing to appeal to their consumers in light of the pandemic, and to explore how online Twitter uses responded to this campaign. It has been predicted that Covid-19 has increased precarious work conditions amongst all workers, but particularly those deemed essential (Davenport et al., 2020; Kniffin et al., 2020). Thus, it is important to understand how these essential workers are represented in the online media by their employers vis-à-vis how they are treated, and to reflect on how effective these representations may be in shaping public opinion.

LITERATURE REVIEW

The literature review begins by providing a history of allegations of mistreatment against Amazon (2.1), followed by an overview of companies who have allegedly treated their companies unfairly and the repercussions of this (2.2), and an explanation of the term 'woke washing'. Finally, this section finishes

by reviewing some crisis management theories and strategies to shed light on how organisations communicate through times of crisis (2.3)

Allegations of employee mistreatment against Amazon

It's important to recognise that even before the Covid-19 pandemic hit, Amazon had a history of mistreating their employees. In 2011, Amazon forced workers in their Allentown warehouse to work without ventilation when the heat index reached 102 degrees Fahrenheit inside the building. Management would not let them open the windows or doors, citing concerns of theft. 15 workers collapsed as a result of the situation (Verrone, 2019). A 2013 investigation revealed that Amazon tracked the movements of their staff with GPS trackers, routinely gave staff one 30 minute break across a ten hour shift, and laid off temporary staff to avoid giving them the same benefits and rights as permanent workers (4 News, 2013). In addition, an employee attempted suicide in 2016 by jumping off Amazon's Seattle headquarters, after sending an email to their co-workers which criticised the company's treatment of employees (Musumeci, 2016). In 2018, *Business Insider* spoke with 31 current or recently employed delivery drivers and found a number of alleged abuses from their employer, including a lack of overtime pay, missing wages, intimidation, and favouritism. The drivers felt they were under severe time constraints, causing them to drive at dangerously high speeds, ignore stop signs, and urinate in water bottles (Peterson, 2018). In June 2020, Amazon delivery drivers living in Canada launch a $200 million class action against the company, claiming for unpaid wages (Mojtehedzadeh, 2020). These examples illustrate that Amazon's mistreatment of their employees has been long standing and has impacted workers worldwide.

Public awareness of Amazon's working conditions grew considerably in light of Covid-19. The retailer was accused of lacking adequate deep cleaning after there were confirmed COVID-19 cases at its fulfilment centres (Sainato, 2020); failing to provide personal protective equipment, hand sanitiser and socially distanced workspaces (Pound, 2020); and a lack of paid sick leave (Woodward, 2020). Some Amazon employees shared their stories of mistreatment with the news media, arguing the company prioritises productivity targets over employee safety. In March 2020, Amazon confirmed that two workers at the Staten Island Fulfillment centre had tested positive to Covid-19. However, employees at the warehouse alleged this number was understated, and that the number of infected employees was actually ten (Palmer, 2020a). In responses, some workers at the centre walked out on the job to protest the lack of protective measures for Amazon workers. The organiser of this protest, Chris Smalls, was subsequently laid off for what Amazon said was "violating social distancing guidelines" (Palmer, 2020a). However Smalls, and other employees, believed he was fired instead for organising the strike (Evelyn, 2021). The retailer also came under scrutiny in June 2020, after Amazon decided to abruptly end hazard pay – which was an increase in employee wages of an extra $2 per hour that it began offering at the start of the pandemic. The payment was discontinued, even though the pandemic impacts and the number of Covid infections, were continuing to increase (Sherman, 2021).

Amazon has managed to keep unions out of its operations since its founding in 1994 (Palmer, 2020b). However, growing media attention which centred on employee mistreatment meant that calls for unionisation at Amazon gained public traction. In response, Amazon posted a job listing for two intelligence analysts who could monitor "labor organizing threats" and other sensitive topics on their website (Palmer, 2020b). A Vice report also found that the retailer's HR department was monitoring Facebook groups used by drivers as a means of tracking planned strikes and organising activity (Gurley & Cox, 2020). This has led many to allege that Amazon is spying on their workers as a means of hindering the unionisation of their employees.

As the above discussion implies, Amazon as a company received a lot of negative media attention and public scrutiny throughout the pandemic specifically. Moreover, Amazon launched a series of advertisements on both traditional and social media which portrayed the company as being deeply concerned with the safety of its 'retail heroes', as well as highlighting the Covid-safe practices enforced within the company's facilities. However, little is known about how Amazon framed its treatment and relationship with its employees across this campaign, and how the public responded to it. This is exactly what the current paper seeks to address.

The Mistreatment of Employees

There are a number of instances of employees being mistreated by their managers and their organisation more broadly. Korean Air, Coke, George Calombaris's MAdE Establishment Group, Coles, Woolworths, Walmart, Tyson Foods, Apple and Nike are just some examples of companies which have had allegations made about them either underpaying their employees, forcing them to work in dangerous conditions, or a combination of both (Kim & Austin, 2020; Parker, 2016). Employee mistreatment has arguably worsened in light of Covid-19, across all industries, with employees being increasingly forced to work into unsafe conditions which lack Covid-safe measures and the retaliation against those workers who speak up (Davenport et al., 2020). McDonalds, Instacart, the Federal Bureau of Prisons, Dillard's and various hospitals worldwide have come under fire for their treatment of frontline workers throughout the pandemic.

Despite the mistreatment of employees gaining public awareness, minimal research has been conducted on how organisations have communicated and framed these allegations to both internal and external audiences. It is important to research the communication of employee mistreatment because consumers are increasingly valuing brands that treat their employees fairly and ethically (Kim & Austin, 2020). For example, a number of studies have found that the mistreatment of employees has a detrimental impact on brand reputation and business performance (Andreu, Casado-Díaz, & Mattila, 2015; Lyon & Cameron, 2004; Sarwar & Muhammad, 2020). Furthermore, this paper is important in that it considers how Amazon's workers are represented on Amazon's twitter vis-à-vis how they are treated. These insights will then be used to reflect on how effective these representations may be in shaping the perceptions of consumers.

Woke washing

In today's business environment, consumers increasingly expect companies to be responsible corporate citizens, in addition to providing high quality goods and services (Diddi & Niehm, 2016; Fuentes-García, Núñez-Tabales, & Veroz-Herradón, 2008; Hess, Rogovsky, & Dunfee, 2002). A 2018 survey revealed that 64% of consumers would reward firms that engage in some kind of social or political activism (Edelman, 2018). As a result, businesses are increasingly seeking to be 'activists', by attempting to encourage social-political change whilst also seeking reputational and economic benefits (Vredenburg et al., 2020). Brand activist initiatives are typically run in conjunction to Corporate Social Responsibility (CSR) campaigns, through which actions and policies which take into account the triple bottom line are implemented (Aguinis, 2011). Some research has suggested that CSR efforts are positively correlated with an increase in profit (Goering, 2014; Lin, Yang, & Liou, 2009; Orlitzky, 2008; Porter & Miles, 2013) and brand reputation (Brammer & Pavelin, 2006; Navarro, 1988; Waddock & Graves, 1997).

With these economic incentives, however, also comes the potential for brands to act inauthentically to obtain greater profits. Sobande (2019) and Vredenburg et al. (2018) discuss 'woke washing', which

refers to inauthentic brand activism whereby a brand's practices do not align with their messaging. A prominent example of this is Nike's decision to make Colin Kaepernick the face of their 30th anniversary advertising campaign (Duarte, 2020). At face value this was a meaningful stand for racial justice, but public records indicate that in 2019 less than 10% of Nike's 300-plus vice-presidents worldwide were black (Nike, 2020). The bronze Fearless Girl sculpture constitutes another example of a brand's image projection not aligning with its internal practices. The statue, located in New York, was commissioned by State Street Global Advisors as a means of promoting an index fund comprised of companies with a high percentage of female leaders. However, State Street was later accused of underpaying its female employees (Mahdawi, 2018).

The growing use of social media to promote CSR initiatives and ethical conduct makes it easier for brands to capitalise off current social issues with the sole goal of creating profits. This is prevalent in the COVID-19 pandemic, where some businesses have sought to profit from the crisis by inflating prices or making misleading claims about products and services (He & Harris, 2020). It is reasonable to suggest that Amazon has engaged in woke washing, given their traditional and social media advertising campaigns at the surface level represent the organisation as being deeply concerned with the safety of its staff and the enforcement of Covid-safe practices within its facilities. The goal of this study is to examine exactly how Amazon has represented its staff and their safety on Twitter.

A number of studies (Christensen, 1995; Hughes, 2013; Stohl, 1995) have found that external communications, and advertising specifically, has an important influence on both internal and external audiences. Moreover, advertising campaigns can influence and be meaningful to the perceptions, attitudes and behaviours of employees, and the extent to which they identify with their organisation's brand. In this sense, if Amazon is engaging in woke washing and is promoting itself as being concerned with employee safety, this has a direct influence on the beliefs and attitudes of the organisation's workers. This, in addition to employees such as Chris Smalls being fired for standing up for better working conditions, may coerce employees into remaining silent and stop them from sharing their experiences even internally – further contributing to the cycle of employee mistreatment.

Crisis Management Strategies

The negative media attention that Amazon has experienced throughout the last year, constitutes a form of crisis. According to Koster and Politis-Norton (2004), a crisis is a major, abrupt event – which is sometimes unexpected – that has potentially negative consequences for an organisation, its employees, financial situation and brand reputation. A section of the public relations and marketing literature examines crisis management specifically, whereby strategies and approaches are developed to combat crises and to lessen the damage inflicted by a crisis (Coombs, 2014). Furthermore, there are a number of different scholarly approaches to crisis management, including Image Restoration Theory (Benoit, 2008), Attribution Theory (Coombs, 2007a) and Contingency Theory (Pang, Jin, & Cameron, 2010).

One of the most referred to crisis management theories in the literature is Situational Crisis Communication Theory (SCCT), which holds the tenet that the way in which organisations respond to a crisis depends on how the public attributes responsibility for that crisis (Zamoum & Gorpe, 2018). If an organisation perceives the crisis to be intentional, they will attribute more responsibility to the organisation. Moreover, SCCT states that the more the organisation is perceived as being responsible for the crisis, the more the organisation should utilise strategies that acknowledge responsibility for the crisis and demonstrate concern for the victims (Bradford & Garrett, 1995; Coombs, 1995, 2014). SCCT's crisis communication strategies include (Coombs, 2007b):

- Denying the crisis (attack the accuser, denial, using a scapegoat)
- Diminishing the crisis (excuses, justification)
- Rebuilding after the crisis (compensation, apologise)

In addition, the SCCT model states that throughout the crisis, the organisation should remind stakeholders of the organisation's past good works and/or remind stakeholders that the organisation is a victim of the crisis too (Coombs, 2007b)

Whilst the current paper takes a more empirical and inductive approach to the way Amazon has communicated about its treatment of employees, it will use the SCCT framework as a means of evaluating how Amazon has portrayed the crisis to stakeholders as well as discussing ways this could have been improved.

METHODOLOGY

The aim of this paper is to investigate how Amazon uses frames and refers to their employees on Twitter and to investigate users' response to it. The term 'frame', here, refers to the way in which a particular event or issue, which in his case is employee relations, is presented by Amazon. Frames allow audiences to interpret new information by sorting it into recognised categories, consisting of an "interpretative structure that sets particular events within a broader context" (Norris, 1995). Frames emphasize specific aspects of reality, meaning particular attributions, evaluations or decisions are assigned to recipients (Scheufele, 2004). Moreover, the goal of adopting a frame analysis in this paper is to consider what aspects of employee relations do Amazon highlight in their Twitter posts, and how do these aspects seek to create a certain representation of these relationships. As there is little research on how organisations have specifically adopted frames when communicating to audiences throughout crises, this paper adopts an inductive approach to coding. In this sense, the study examines frames in Amazon's tweets and then seeks to make sense of those patterns (Simmons et al., 2011).

In terms of examining how users responded to Amazon's posts, this paper also conducts a thematic analysis of tweet replies. Thematic analysis is a qualitative research method which identifies, analyses, explains and organises themes within a given data set (Clarke & Braun, 2014). A thematic analysis allows this study to identify what the common themes and topics users are talking about when they are replying to Amazon, thereby providing insight on how audiences perceive Amazon's tweets regarding their employees. Again, this paper adopts an inductive approach to coding, identifying themes intrinsic to the data. It is timely to note that when undertaking both the thematic analysis on user replies and the frame analysis on Amazon's posts, it was possible for the content to contain more than one frame or theme.

This study collects the posts tweeted by Amazon on Twitter between 1st April to 30th December 2020 which mentioned their employees specifically, in addition to users' responses to these posts. The purpose of specifically collecting posts which referred to employees is to fulfill this paper's aim of investigating how Amazon framed his employees on social media. When collecting users tweet replies to these posts, the tweet had to contain at least two words (unless it was a hashtag) to be included in the analysis. In addition, tweet replies that just contained links, gifs, images or other users' names without any additional text were excluded from the analysis. In sum, this study collected a total of 40 tweets from Amazon, and 2843 tweet replies.

A qualitative content analysis was conducted on Amazon's Twitter posts to investigate the frames the company used in representing itself and its employees whilst allegations of employee mistreatment accrued in the media. A qualitative content analysis is also conducted on users' Twitter replies to understand how online audiences responded to the campaign through specific themes.

RESULTS

This section beings by analysing the results of the content analysis of the frames used by Amazon in their tweets (3.1). This is then followed by an analysis of user tweets which were responding to Amazon's posts (3.2).

Results of Content Analysis of Amazon's Tweets

Firstly, this paper conducted a qualitative content analysis of Amazon's Twitter posts to explore how Amazon framed and talked about its employees in its posts. The results of this content analysis are shown in Table 1.

Table 1. Results of qualitative content analysis of Amazon's frames on Twitter.

Frame	Number of posts	Percentage of all posts collected (%)
Amazon employees are family	19	47.5
Our employees are appreciated	13	32.5
Amazon employees are heroes	7	17.5
Amazon is a hero	7	17.5
Amazon is a great place to work	5	12.5

The 'Amazon Employees are Family' Frame

The most dominant frame in Amazon's tweets concerning their employees was 'Amazon employees are family', which was found in 47.5% of the 40 posts analysed in this paper. Amazon exhibited a strong tendency to portray its employees as being more than workers, but rather members of their family unit. This is illustrated in the following tweets:

> "*We feel like a family together*. It's more than just a great job for these @Amazon
>
> employees. Hear their stories [Down pointing backhand index] #teamthatdelivers https://amzn.to/3pQm01D" (Amazon, 2020h)
>
> "The Amazon family couldn't be more proud of this everyday hero [Red heart] Thank you, Sean [Raising hands]"
>
> "Janelle's a proud Area Manager and even prouder mom. She knows how important staying safe for your family is and has been taking care of her work family too—shipping millions of masks to teams across our network. See more Amazon stories on our blog. https://amzn.to/3cN1zL5" (Amazon, 2020c)

In these tweets, Amazon indirectly denies any allegations of employee mistreatment by suggesting that their workers a strong rapport with upper-level management. In doing so, this frame aims to discourage any negative beliefs audiences may have about Amazon and their treatment of staff.

In addition, the 'Amazon employees are family' frame was activated through the use of 'motherly' keywords such as 'caring' (e.g. "We devote enormous time and resources to caring for our people" (Amazon News, 2020)), 'protecting' (e.g. "see all of the ways we're working to protect our people" (Amazon News, 2020)), 'helping' (e.g. "Christine's an Amazon Seasonal Sortation Associate and mom with a passion for helping people. She's proud to look out for her team's health…" (Amazon, 2020a)) and 'supporting' (e.g. "...@AmazonNews has daily updates about the actions we're taking to support our people" (Amazon, 2020i)). Through these words, Amazon's tweets represent the organisation as having a motherly and nurturing role over the health and wellbeing of its employees. In doing so, 'Amazon employees are family' minimises the experiences of employees who may have a differing experience to what the frame represents.

The 'Our Employees are Appreciated' Frame

The second most common frame used by Amazon was 'Our employees are appreciated'. Tweets containing this frame often highlighted the achievements of their staff, and how 'proud' the organisation is of these workers. In these tweets, employees were appreciated for doing their jobs, as well as other unique accomplishments such as saving a dog's life and returning a lost goat home whilst making deliveries, and gifting a cancer patient flowers and a card in addition to their package. This frame aims to deflect attention from the employee mistreatment allegations. According to Martin (2021), a central issue with organisations publicly expressing appreciation for their employees is that it implies that the contributions of employees are a form of beneficence, thus masking the employer's power over the employee. Gratitude implies that someone has benefitted from the person they are grateful to. Portraying employee's workplace contributions as benefits is therefore problematic as benefits require downplaying management's power over the employee (Martin, 2021). Moreover, in some instances, the 'our employees are appreciated' frame normalises an oppressive power relationship between Amazon's management and their workers.

The 'Amazon employees are heroes' Frame

In their tweets, Amazon exhibited the tendency to emphasise the role of their frontline workers in a society drastically impacted by Covid-19, and how much society depends on them. By highlighting the dependence of society on their front-line workers, this frame portrays Amazon's workers as heroes whilst also normalising the exposure of their employees to unsafe working conditions and risks. Consider the following tweets made by Amazon:

> "Today's visits by our founder and CEO @JeffBezos to say thank you to Amazon fulfillment center and @WholeFoods employees. We're all incredibly proud of the thousands of our colleagues working on the front lines to get critical goods to people everywhere during this crisis" (Amazon, 2020f)
>
> "Rising to the challenge is what our people do, like Kent, an Area Manager working to support his young son. He knows people depend on him for their groceries and products, and he's proud to deliver for them. See more employee stories on our blog. https://amzn.to/3e2FCsy" (Amazon, 2020e)

The above examples present Amazon's frontline workers in a heroic light by highlighting their role in maintaining a functioning society throughout global lockdowns. This heroic discourse thereby portrays Amazon's employees as undertaking a moral and sacrificial act by working in dangerous conditions. In

doing so, the frame makes the unacceptable - such as having inadequate access to Covid-safe measures, a lack of paid sick leave and approximately 20,000 Covid-19 cases among Amazon workers (BBC News, 2020b) – more acceptable to audiences. In this sense, the 'Amazon employees are heroes' frame conceals the dangerous and unfair conditions many Amazon workers face. The use of 'heroic' rhetoric to normalise unsafe working conditions has also been used in public discourse surrounding the role of nurses through Covid-19 pandemic (Mohammed et al., 2021) and the SARS crisis (Hall et al., 2003).

The 'Amazon is a 'hero' frame

This frame was used to suggest that Amazon had gone above and beyond what was expected of them for their employees. Tweets containing this frame often portrayed Amazon as a hero for fulfilling their basic role as an employer – maintaining the safety of their employees. Consider the following tweets by Amazon:

> "We devote enormous time and resources to caring for our people — it's our #1 priority. From masks to physical distancing to temperature checks, see all of the ways we're working to protect our people" (Amazon News, 2020)
>
> "We don't just think big, we do big. We've shipped over 100 million masks to our network and we're spending $4 billion to keep employees safe and get people what they need. We'll never stop doing our part: https://amzn.to/2SWqYuI" (Amazon, 2020g)
>
> "Millions of masks, gloves, and cleaning supplies add up to one thing: Safety. Our people's health comes first, and we've been working around the clock to get them what they need to stay safe" (Amazon, 2020d)

By portraying themselves as heroes, Amazon seeks to deny any allegations of employee mistreatment by trying to prove to audiences that they take the health and safety of their workers seriously. Similar to the 'Amazon is a family' frame, 'Amazon is a hero' denies the experience of any employees who have been mistreated or work in unsafe conditions, and as such is a form of oppression from management over front-line workers.

The 'Amazon is a great place to work' frame

This frame was used to suggest that Amazon has a great work culture and that employees enjoy working there. It was also used by Amazon to appeal and communicate to potential employees, with the frame often being utilised in tweets promoting recruitment at the organisation. Tweets containing this frame deflected attention away from employee mistreatment allegations, and focused on the opportunities Amazon promoted themselves as offering. Consider the following examples:

> "Employees like Jerome make sure new hires are set up for success, whether long-term or short-term. For more info, visit our blog. https://amzn.to/3fOWHYA" (Amazon, 2020b)
>
> "We're getting excited for #AmazonCareerDay on September 16th! Watch Amazon Scout help kick off the event, offering this new employee a special welcome to the company. Handshake Learn more here: https://amzn.to/33nhsFc" (Amazon, 2020j)

Results of Content Analysis of replies to Amazon Tweets

In addition, this paper also combined a thematic analysis with a content analysis to identify the common themes Twitter users were talking about when responding to Amazon's posts about employees. The results of this are shown in Table 2.

Table 2. Results of qualitative content analysis on user replies.

Theme	Number of replies	Percentage of all tweets collected (%)
Customer Complaints	953	33.64
Customer enquiries and questions	448	15.81
Mistreatment of employees	408	14.40
Low wages and lack of hazard pay	250	8.82
Praising Amazon	197	6.93
Concerns of employee safety	187	6.58
Other (could not be categorised as any theme)	170	5.98
Calling out Amazon for woke washing	131	4.61
Wanting Amazon to open to Pakistan	112	3.94
Customer suggestions	96	3.38
politics and 2020 election	91	3.20
Concerns about PlayStation 5	78	2.74
Increased wealth of Jeff Bezos	73	2.57
Complaints made by employees	60	2.11
Concerns about products	48	1.69
Promoting the unionisation of Amazon employees	40	1.41
Concerns of price gouging	38	1.34
Threats to boycott Amazon	37	1.30
Concerns of organisational racism	30	1.06
Wanting a job with Amazon	28	0.98
Concerns of Amazon avoiding Tax	24	0.84
Agreeing with Amazon's actions	20	0.70
Employees praising Amazon	17	0.60
Critiques of Capitalism	5	0.18

A number of themes were identified in the corpus. Some of these related to Amazon's treatment of their employees, whilst many did not. As this paper is concerned with how Twitter users perceived and reacted to Amazon's tweets regarding their employees, this section will only provide in-depth analysis on themes related to this topic.

In general, the results of this paper indicate that a considerable amount of Twitter users were concerned with the treatment of Amazon workers by their employer. 408 tweet replies contained the 'mistreatment of employees' theme, 250 contained 'Low wages and lack of hazard pay', 187 contained 'Concerns of employee safety', 131 contained 'Calling out Amazon for woke washing', 73 contained 'Increased wealth of Jeff Bezos' and 40 contained 'Promoting the unionisation of Amazon employees'. On top of this, 60 tweets were made by people claiming to be employed by Amazon, and were complaining about their treatment. This compares to 197 tweets which responded positively to Amazon's posts and praised the organisation, 20 tweets which actively sided with Amazon's treatment of their employees and 17 tweets which were written by users who claimed to be employed by Amazon and were praising Amazon's treatment of their employees.

Under the 'mistreatment of employees' theme, tweets did not specifically mention if they were concerned with the conditions Amazon's frontline workers face or a lack of pay, but rather they expressed a concern for the wellbeing of the employees at a general level. For example:

> "You keep putting out these Amazon is awesome & cares about their employees but I can't believe them & I bet if I asked any Amazon employee-especially in the warehouses- they'd confirm these commercials are pure fiction. Spend your PR money on your actual employees"
>
> "And he only get a palm in the back, like the rest of the employees ..while they make their boss richer and richer risking their life's in this pandemic ..bravoo".

Separate to this was tweets which were concerned with Amazon's Low wages and their decision to stop hazard pay specifically. These tweets were often highly emotional and angry in nature:

> "amazon associates would be a lot better off if you didn't take away our bonuses. It makes me sick knowing that tier 4 & managers still get a bonus. Tier 1 gets the occasional swag. Warehouses are STILL hazardous, yet, our hazard pay expires in a couple days. Soooo greedy".

This emotion-laden discourse was also common amongst tweets concerned with employee safety:

> "7 @amazon employees dead. No transparency into how many are sick & Amazon will end bonus pay for workers while they make @JeffBezos the first $trillionaire. This is why we #boycottAmazon @BobFergusonAG"

Importantly, a large number of tweets seemed to call out Amazon for false advertising or 'woke washing' when the company posted that their employees were treated fairly:

> "Your news propaganda piece is more reason to distrust your company. It makes it seem like you want to control the narrative, which comes across as a the lie that it is"
>
> "Your commercials don't make me think you treat employees better. @Amazon is about greed and Bezos becoming a trillionaire. #greedoverintegrity #greedoveremployees #greedoverpeople"

The amount of Twitter replies containing the 'Calling out Amazon for woke washing' theme (4.62% of all tweets collected) suggest that a considerable portion of online users actively accused Amazon of engaging in some form of woke washing. 'Complaints made by employees' was commonly used with the woke washing theme, in that many employees attacked Amazon for lying about working conditions in their Twitter posts:

> "STOP LYING, IM A DRIVER, WE STILL DONT GET TEMPERATURE CHECKS, WE STILL HAVEN'T FOUND OUT ABOUT INFECTED WORKERS AT DSP1. PROFITS OVER SAFETY"

73 tweets also expressed their dismay for Amazon's treatment of workers, by highlighting CEO Jeff Bezos' growing net worth. Tweet replies containing the 'Increased wealth of Jeff Bezos' argued that Amazon can afford to pay their employees more and treat their workers better, given that Bezos is currently the richest person alive. For example:

> "Amazon continues to treat its warehouse workers like expendable pawns, by not paying sick leave and implementing job protection guarantees. All this while Bezos adds billions to his personal bank account"
>
> "Jeff Bezos to become world's first trillionaire from the profits he made on the back of a pandemic while we are about to enter a global recession and Amazon workers lose hazard payments? This is why capitalism is wrong, such wealth should never be in the hands of an individual"

Moreover, tweets containing this theme often also contained the 'Low wages and lack of hazard pay' theme in that they were highly critical of Amazon's decision to stop hazard pay, given Bezos' current net worth.

1.41% of tweet replies also either encouraged Amazon's workers to form a union and/or stressed the importance of unionisation – particularly in regard to ensuring employees are paid fairly:

> "Still not a living wage. Amazon should be unionized" and "You are union-breaking employers that don't allow loo breaks and force workers to meet impossible standards! What lies you tell. Bezos is cruel to the workers that make him excessively rich.
>
> #BoycottAmazon #MakeAmazonPay #BuyNothingDay #BlackFriday2020"

Tweets containing the 'Promoting the unionisation of Amazon employees' theme also criticised Amazon for preventing employees from forming a union:

> "Such an empty PR vid when it is well known you are (and have been) actively trying to prevent them from unionizing" and "How much money do you spend each year on union busting?".

It is important to recognise that a number of Twitter replies responded positively to Amazon's tweets about their employees (6.95% of total tweets), siding with Amazon's treatment of their employees (0.71% of total tweets), as well as a number of employees coming forward to deny the allegations circulating about employee mistreatment (0.6% of total tweets). When Amazon published a gratifying tweet about an employee who was previously a veteran, one Twitter user responded "good job @amazon Great support for our veterans". Similarly, when Amazon posted a video of Jeff Bezos visiting one of their warehouses to thank the employees for their hard work, a user replied "Nicely maintained fitness Jeff [Thumbs up] Nice motivation booster to employees [Flag of United States]".

In addition, some tweets sided with Amazon and dismissed any allegations of employee mistreatment as necessary or deserved. For example: "for all this those complaining, if you don't think they pay enough, don't work there [Man shrugging]" and "No complaint from me for firing insubordinate employees. They knew what they signed up for. Get another job if you can't follow the policies. Now, if my order doesn't come on time I might change my mind. Just kidding!". One user justified Amazon's low wages by claiming it would cost employees their jobs: "Raising the minimum wage only increases the costs of goods. It also forces companies like Amazon and McDonald's to look for automation solutions. This means the cost of goods will be higher with a decrease in available jobs!".

Finally, a smaller number of tweets were written by people claiming to work for Amazon and praised the organisation for the way they treat their staff. When one Twitter user referred to allegations that Amazon workers were forced to 'pee in water bottles' as they were not allowed restroom breaks, another user replied: "What absolute BS lol You can take a restroom break whenever you want".

Likewise, when a user claimed that Amazon did not pay their staff more because they 'lack human decency', another user replied:

> 'I have been working here at Amazon and I have had no complaints. Is there anything in specific that you want to bring up? Id be more than happy to discuss my experience here".

Furthermore, tweets embodying the 'Employees praising Amazon' theme were often responding to other users who were criticising Amazon's practices. However, it is important to be cautious with these more positive tweets about Amazon, given that the retailer is known for paying users to tweet positively about Amazon (BBC News, 2021).

DISCUSSION

The results of this paper have shed light on both how Amazon sought to portray its own treatment of employees, as well as how Twitter users perceived and responded to this portrayal. From the perspective of Amazon, it is clear through this paper's analysis of frames used by the organisation on Twitter, that Amazon was actively denying and avoiding allegations of employee mistreatment. SCCT claims that denial crisis management strategies involve claiming that no crisis exists or declaring that the organisation is not responsible for it (Claeys, Cauberghe, & Vyncke, 2010). At least on Twitter, Amazon failed to address the widespread allegations on employee mistreatment, making the organisation come across as arrogant, uncaring and unsympathetic (Davies, 2005).

The choice to deny the allegations is particularly risky in Amazon's case, given there are so many people involved who may be able to provide evidence to counter Amazon's narrative. Unlike other types of crises, such as a financial crisis, where the information surrounding the event is typically concentrated amongst a small number of people, Amazon's mistreatment allegations concerns the company as a whole and the experiences of every one of Amazon's 1,298,000 employees can either reinforce the allegations or dispute them (Statista, 2021). Given this large number of people involved in the crisis, it can be – and has been – difficult for Amazon to control the denial narrative. For example, Amazon actively denied allegations that their workers were forced to pee in water bottles due to harsh quotas, tweeting in a reply to US Representative Mark Pocan:

> "1/2 You don't really believe the peeing in bottles thing, do you? If that were true, nobody would work for us. The truth is that we have over a million incredible employees around the world who are proud of what they do, and have great wages and health care from day one" (Amazon News, 2021).

However, in March 2021, *The Intercept* obtained a confidential document from one of Amazon's employees, in which Amazon acknowledged that the practice had occurred. In addition, given that SCCT states that the more the organisation is perceived as being responsible for the crisis, the more the organisation should acknowledge responsibility for the crisis (Bradford & Garrett, 1995; Coombs, 1995, 2014), it may have been better from a branding perspective for Amazon to accept some responsibility for the allegations, given that it is difficult for the retailer to blame anyone else with this specific crisis.

Moreover, Amazon's could have used SCCT's excuse strategy, through which they could acknowledge the crisis but deny any intent to do harm (Coombs, 2007b), or by apologising and undertaking an official investigation into the mistreatment allegations. Even in April 2021, when Amazon announced

they were giving over 500,000 employees a pay increase between at 50 cents and $3 an hour, the company failed to acknowledge allegations thar staff have been underpaid or that they were apologetic for dropping hazard pay the previous year:

> "More than 500,000 people will see an increase between at least 50 cents and $3 an hour, which is an investment of over $1 billion in incremental pay for these employees. This is on top of our already industry-leading starting wage of at least $15 an hour and the more than $2.5 billion that we invested last year in additional bonuses and incentives for front-line teams. These jobs come with a range of great benefits, like medical, dental, and vision coverage, parental leave, ways to save for the future, and opportunities for career advancement—all in a safe and inclusive environment that's been ranked among the best workplaces in the world" (Henry, 2021)

Given, as this study has shown, Amazon's narrative that the organisation is a great and fair place to work has been contested by audiences, examples like the one above illustrate how the company's denial makes them come across out of touch and arrogant. This paper's analysis of how Amazon utilised frames when tweeting about their employees show how the company used Twitter to maintain oppressive relationships, to normalise the dangerous and unsafe working conditions that many of Amazon's employees face, and to minimise and silence the experiences of employees who may have had different working experiences to the one Amazon was projecting. As discussed below, a central way to counteract Amazon's narrative is to form a collective labour movement on social media where employees, even anonymously, can share their experiences and place further pressure on the company to change their practices.

Whilst social media makes it easier for organisations to engage in marketing and communications, with it also comes the potential for a loss of control over the message, and the dilution of common frames and identities as other users engage and interact with the content (Dumitrica & Felt, 2019). The results of this study's content analysis of users' replies to Amazon's posts show that people primarily use Amazon's Twitter page as a platform to make complaints about their own personal experiences with the company, or to make customer enquiries. This paper has also shown that a consider proportion of users used Twitter to call out Amazon for what they perceive is injustice and for promoting an image of itself that does not necessarily align with its practices – 'woke washing'. However, there is still a lot of room for social media to place pressure on Amazon to treat its employees better.

Despite Amazon increasing the wages of 500,000 workers, many employees still feel they are being treated unfairly (Fickenscher, 2021; Greenhouse, 2021b). In the lead up to a unionisation vote in Alabama, whereby workers ended up failing to unionise, Amazon spread anti-union propaganda by creating an anti-union website, requiring all workers to attend group meetings where the company promoted its anti-union views, sending anti-union text messages and placing anti-union posters in the bathroom (Greenhouse, 2021a). Furthermore, social media offers a platform for employees and wider society to encourage Amazon to treat their workers better. Whilst this paper has shown that some Twitter users have actively called out Amazon for woke washing and have used social media to express their concerns of employee mistreatment, this is not enough to force Amazon to provide better conditions for their workers. Given that in the US, companies have the right to ban non-employee union organizers from their property, moving forward – specifically if Amazon employees want to unionise – social media is an important tool for workers to mobilise and place pressure on Amazon. Labour campaigns such as 'Our Walmart' and 'Fight for 15' have shown how social media can be used to create a collective identity amongst dispersed workers, enhance mobilisation, allows workers to share their experience of injustices within the workplace (even anonymously), and can be used as a

form of support for workers. Moreover, if Amazon staff, or an independent organisation working on their behalf, were to form a collective labour movement online, it would allow these employees to amplify the offline actions of the movement, create activist networks, and support the emergence of new activist activities (Pasquier & Wood, 2018). Whilst there is a Twitter page titled 'Amazon Labor Union' which claims to "worker-led movement to put power back in the hands of Amazon employees and establish our right to negotiate for a better, safer, and more equitable workplace" (Amazon Labor Union, 2020), this page was only formed in September 2020 and did not engage with any of Amazon's posts analysed in the study. There is therefore still scope for Amazon Labour Union, and similar accounts to interact more with Amazon on social media, contest their narrative, and mobilise public support for better working conditions.

CONCLUSION

The goal of this paper was to consider how, in light of accruing allegations of employee mistreatment, Amazon framed itself and its employees on Twitter, and ultimately engaged in woke washing. In addition, this paper also was interested in examining how Twitter users understood these frames, and whether they believed Amazon's content or whether they still had concerns over the retailer's alleged mistreatment. In summary, this paper found that Amazon used frames which maintained oppressive relationships, normalised dangerous and unsafe working conditions, and silenced the experiences of employees. In terms of perceptions, the thematic analysis of user comments revealed that whilst a considerable number of comments actively called out Amazon for woke washing on social media or still expressed concerns for alleged employee mistreatment, there were not enough of these comments to put pressure on Amazon into changing. Moreover, this paper discussed the scope for more organised and collective efforts to use social media to create change in Amazon's working conditions and employee wellbeing.

There are a number of limitations to this study. Firstly, this paper only collected posts from Amazon that referred to their employees. However, to provide a more holistic understanding of how users perceived Amazon's treatment of their employees, it would be useful to have collected other posts which did not explicitly refer to or cite their staff. Secondly, this paper is very qualitative in nature, and thus subjective. Future research could consider repeating the methodology in a different time period, to corroborate the results of this paper. Future research could also explore how Amazon has represented itself on other platforms, like television and on its blogs, to make comparisons on how the company may seek to represent itself to different audiences.

KEYWORDS: woke washing, social media, marketing, Covid-19, Amazon.

REFERENCES

4 News. (2013, 1 August 2013). Anger at Amazon working conditions. Retrieved from https://www.channel4.com/news/anger-at-amazon-working-conditions

Aguinis, H. (2011). Organizational responsibility: Doing good and doing well.

Amazon (2020a). Christine's an Amazon Seasonal Sortation Associate and mom with a passion for helping people. She's proud to look out for her team's health through daily temp checks and her care for others is an inspiration to her son. See more Amazon stories on our blog. https://amzn.to/37viwsw [Twitter]. Retrieved from https://twitter.com/amazon/status/1272876390237573121

Amazon [@Amazon]. (2020b). Employees like Jerome make sure new hires are set up for success, whether long-term or short-term. For more info, visit our blog. https://amzn.to/3fOWHYA [Twitter]. Retrieved from https://twitter.com/amazon/status/1261279843070640129

Amazon (2020c). Janelle's a proud Area Manager and even prouder mom. She knows how important staying safe for your family is and has been taking care of her work family too—shipping millions of masks to teams across our network. See more Amazon stories on our blog. https://amzn.to/3cN1zL5 [Twitter]. Retrieved from https://twitter.com/amazon/status/1270702046551965697

Amazon (2020d). Millions of masks, gloves, and cleaning supplies add up to one thing: Safety. Our people's health comes first, and we've been working around the clock to get them what they need to stay safe.

For more safety info, visit our blog: https://amzn.to/3f0e9sB [Twitter]. Retrieved from https://twitter.com/amazon/status/1255481739520151554

Amazon (2020e). Rising to the challenge is what our people do, like Kent, an Area Manager working to support his young son. He knows people depend on him for their groceries and products, and he's proud to deliver for them. See more employee stories on our blog. https://amzn.to/3e2FCsy [Twitter]. Retrieved from https://twitter.com/amazon/status/1265266163665252359

Amazon (2020f). Today's visits by our founder and CEO @JeffBezos to say thank you to Amazon fulfillment center and @WholeFoods employees. We're all incredibly proud of the thousands of our colleagues working on the front lines to get critical goods to people everywhere during this crisis. [Twitter]. Retrieved from https://twitter.com/amazon/status/1248094415752749059

Amazon (2020g). We don't just think big, we do big. We've shipped over 100 million masks to our network and we're spending $4 billion to keep employees safe and get people what they need. We'll never stop doing our part: https://amzn.to/2SWqYuI [Twitter]. Retrieved from https://twitter.com/amazon/status/1260555121521504256

Amazon (2020h). "We feel like a family together." It's more than just a great job for these @Amazon employees. Hear their stories Down pointing backhand index #teamthatdelivers https://amzn.to/3pQm01D [Twitter]. Retrieved from https://twitter.com/amazon/status/1330967808252014604

Amazon (2020i). We're continuing to closely monitor the impact of COVID-19. @AmazonNews has daily updates about the actions we're taking to support our people, customers and the community during this pandemic. Retrieved from https://twitter.com/amazon/status/1250136634819043329

Amazon [@Amazon]. (2020j). We're getting excited for #AmazonCareerDay on September 16th! Watch Amazon Scout help kick off the event, offering this new employee a special welcome to the company. Handshake Learn more here: https://amzn.to/33nhsFc [Twitter]. Retrieved from https://twitter.com/amazon/status/1305569513274998785

Amazon Labor Union [@amazonlabor]. (2020). [Twitter]. Retrieved from https://twitter.com/amazonlabor?lang=en

Amazon News (2020). We devote enormous time and resources to caring for our people — it's our #1 priority. From masks to physical distancing to temperature checks, see all of the ways we're working to protect our people. #WorldSafetyDay https://amzn.to/2SeKimG [Twitter]. Retrieved from https://twitter.com/amazonnews/status/1255129641242824704

Amazon News [@amazonnews]. (2021). 1/2 You don't really believe the peeing in bottles thing, do you? If that were true, nobody would work for us. The truth is that we have over a million incredible

employees around the world who are proud of what they do, and have great wages and health care from day one. Retrieved from https://twitter.com/amazonnews/status/1374911222361956359

Andreu, L., Casado-Díaz, A. B., & Mattila, A. S. (2015). Effects of message appeal and service type in CSR communication strategies. *Journal of Business Research, 68*(7), 1488-1495.

Asian Development Bank. (2020, 15 May 2020). COVID-19 Economic Impact Could Reach $8.8 Trillion Globally — New ADB Report. Retrieved from https://www.adb.org/news/covid-19-economic-impact-could-reach-8-8-trillion-globally-new-adb-report

BBC News. (2020a). Coronavirus: US economy sees sharpest contraction in decades. Retrieved from https://www.bbc.com/news/business-53574953

BBC News. (2020b, 2 October 2020). Nearly 20,000 Covid-19 cases among Amazon workers. Retrieved from https://www.bbc.com/news/business-54381928

BBC News. (2021, 30 March 2021). 'Fake' Amazon workers defend company on Twitter. Retrieved from https://www.bbc.com/news/technology-56581266

Benoit, W. L. (2008). Image restoration theory. *The International Encyclopedia of Communication*.

Bradford, J. L., & Garrett, D. E. (1995). The effectiveness of corporate communicative responses to accusations of unethical behavior. *Journal of business ethics, 14*(11), 875-892.

Brammer, S. J., & Pavelin, S. (2006). Corporate reputation and social performance: The importance of fit. *Journal of Management Studies, 43*(3), 435-455.

Christensen, L. T. (1995). Buffering organizational identity in the marketing culture. *Organization Studies, 16*(4), 651-672.

Claeys, A.-S., Cauberghe, V., & Vyncke, P. (2010). Restoring reputations in times of crisis: An experimental study of the Situational Crisis Communication Theory and the moderating effects of locus of control. *Public Relations Review, 36*(3), 256-262. doi:https://doi.org/10.1016/j.pubrev.2010.05.004

Clarke, V., & Braun, V. (2014). Thematic analysis. In *Encyclopedia of critical psychology* (pp. 1947-1952): Springer.

Coombs, W. T. (1995). Choosing the right words: The development of guidelines for the selection of the “appropriate” crisis-response strategies. *Management communication quarterly, 8*(4), 447-476.

Coombs, W. T. (2007a). Attribution theory as a guide for post-crisis communication research. *Public Relations Review, 33*(2), 135-139.

Coombs, W. T. (2007b). Protecting organization reputations during a crisis: The development and application of situational crisis communication theory. *Corporate Reputation Review, 10*(3), 163-176.

Coombs, W. T. (2014). *Ongoing crisis communication: Planning, managing, and responding*: Sage Publications.

Crane, A., & Matten, D. (2020). COVID-19 and the future of CSR research. *Journal of Management Studies*.

Davenport, C., Bhattarai, A., McGregor, J., & (2020, 1 April 2020). As coronavirus spreads, so do reports of companies mistreating workers.

Davey, J. (2020, 1 December). Topshop owner Arcadia collapses into administration. Retrieved from https://www.smh.com.au/business/companies/topshop-owner-arcadia-collapses-into-administration-20201201-p56jd2.html

Davies, D. (2005). Crisis management: Combating the denial syndrome. *Computer Law & Security Review, 21*(1), 68-73.

Dennis, M. J. (2020). The impact of COVID-19 on the world economy and higher education. *Enrollment Management Report, 24*(9), 3-3. doi:https://doi.org/10.1002/emt.30720

Diddi, S., & Niehm, L. S. (2016). Corporate social responsibility in the retail apparel context: Exploring consumers' personal and normative influences on patronage intentions. *Journal of Marketing Channels, 23*(1-2), 60-76.

Duarte, F. (2020, 13 June 2020). Black Lives Matter: Do companies really support the cause? Retrieved from https://www.bbc.com/worklife/article/20200612-black-lives-matter-do-companies-really-support-the-cause

Dumitrica, D., & Felt, M. (2019). Mediated grassroots collective action: negotiating barriers of digital activism. *Information, communication & society*, 1-17.

Edelman. (2018). Brands Take a Stand 2018. Retrieved from https://www.edelman.com/earned-brand

Evelyn, K. (2021, 1 April 2020). Amazon fires New York worker who led strike over coronavirus concerns. Retrieved from https://www.theguardian.com/us-news/2020/mar/31/amazon-strike-worker-fired-organizing-walkout-chris-smallls

Fickenscher, L. (2021, 10 May 2021). Amazon worker in Alabama claims conditions tougher since union vote. Retrieved from https://nypost.com/2021/05/10/amazon-worker-in-alabama-claims-conditions-tougher-since-union-vote/

Fuentes-García, F. J., Núñez-Tabales, J. M., & Veroz-Herradón, R. (2008). Applicability of corporate social responsibility to human resources management: Perspective from Spain. *Journal of business ethics, 82*(1), 27-44.

Goering, G. E. (2014). The profit-maximizing case for corporate social responsibility in a bilateral monopoly. *Managerial and Decision Economics, 35*(7), 493-499.

Greenhouse, S. (2021a). The union loss at Amazon is another sign big companies have too much power. Retrieved from https://edition.cnn.com/2021/04/23/perspectives/amazon-union-vote-workers/index.html

Greenhouse, S. (2021b, 23 February 2021). 'We deserve more': an Amazon warehouse's high-stakes union drive. Retrieved from https://www.theguardian.com/technology/2021/feb/23/amazon-bessemer-alabama-union

Gurley, L. K., & Cox, J. (2020, 2 September 2020). Inside Amazon's Secret Program to Spy On Workers' Private Facebook Groups. Retrieved from https://www.vice.com/en/article/3azegw/amazon-is-spying-on-its-workers-in-closed-facebook-groups-internal-reports-show

Hall, L. M., Angus, J., Peter, E., O'Brien-Pallas, L., Wynn, F., & Donner, G. (2003). Media portrayal of nurses' perspectives and concerns in the SARS crisis in Toronto. *Journal of Nursing Scholarship, 35*(3), 211-216.

He, H., & Harris, L. (2020). The Impact of Covid-19 Pandemic on Corporate Social Responsibility and Marketing Philosophy. *Journal of Business Research*.

Henry, D. (2021, 28 April 2021). A message from Darcie Henry, Amazon VP of People eXperience and Technology, Worldwide Consumer. Retrieved from https://www.aboutamazon.com/news/operations/a-message-from-darcie-henry-amazon-vp-of-people-experience-and-technology-worldwide-consumer

Hess, D., Rogovsky, N., & Dunfee, T. W. (2002). The next wave of corporate community involvement: Corporate social initiatives. *California Management Review, 44*(2), 110-125.

Hughes, D. E. (2013). This ad's for you: the indirect effect of advertising perceptions on salesperson effort and performance. *Journal of the Academy of Marketing Science, 41*(1), 1-18. doi:10.1007/s11747-011-0293-y

Kim, S., & Austin, L. L. (2020). Employee Mistreatment Crises and Company Perceptions. *International journal of communication, 14*, 21.

Kniffin, K. M., Narayanan, J., Anseel, F., Antonakis, J., Ashford, S. J., Bakker, A. B., . . . Choi, V. K. (2020). COVID-19 and the Workplace: Implications, Issues, and Insights for Future Research and Action.

Koster, M. C., & Politis-Norton, H. (2004). Crisis Management Strategies. *Drug Safety, 27*(8), 603-608. doi:10.2165/00002018-200427080-00011

Lin, C.-H., Yang, H.-L., & Liou, D.-Y. (2009). The impact of corporate social responsibility on financial performance: Evidence from business in Taiwan. *Technology in Society, 31*(1), 56-63.

Lyon, L., & Cameron, G. T. (2004). A relational approach examining the interplay of prior reputation and immediate response to a crisis. *Journal of public relations research, 16*(3), 213-241.

Mahdawi, A. (2018). Woke-washing brands cash in on social justice. It's lazy and hypocritical. Retrieved from https://www.theguardian.com/commentisfree/2018/aug/10/fellow-kids-woke-washing-cynical-alignment-worthy-causes

Martin, A. M. (2021). Against Mother's Day and Employee Appreciation Day and Other Representations of Oppressive Expectations as Opportunities for Excellence and Beneficence. *Pacific Philosophical Quarterly, 102*(1), 126-146.

Mohammed, S., Peter, E., Killackey, T., & Maciver, J. (2021). The "nurse as hero" discourse in the COVID-19 pandemic: A poststructural discourse analysis. *International Journal of Nursing Studies, 117*, 103887. doi:https://doi.org/10.1016/j.ijnurstu.2021.103887

Mojtehedzadeh, S. (2020, 27 June 2020). Amazon delivery drivers in Canada launch $200 million class action claiming unpaid wages. Retrieved from https://www.thestar.com/business/2020/06/26/amazon-delivery-drivers-in-canada-launch-200-million-class-action-claiming-unpaid-wages.html

Musumeci, N. (2016, 29 November 2016). Amazon employee jumps off company building after ranting email to staff. Retrieved from https://nypost.com/2016/11/29/amazon-employee-jumps-off-company-building-after-ranting-email-to-staff/

Navarro, P. (1988). Why do corporations give to charity? *Journal of business*, 65-93.

Neate, R. (2020, 16 April 2020). Amazon reaps $11,000-a-second coronavirus lockdown bonanza. Retrieved from https://www.theguardian.com/technology/2020/apr/15/amazon-lockdown-bonanza-jeff-bezos-fortune-109bn-coronavirus

Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., ... Agha, R. (2020). The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International journal of surgery (London, England), 78*, 185.

Nike. (2020, 10 February 2020). What to Know About NIKE, Inc.'s Latest Impact Report. Retrieved from https://news.nike.com/news/nike-impact-report-fy19

Nine News Australia. (2020, 9 November 2020). Fashion designer Alice McCall driven into voluntary administration by coronavirus. Retrieved from https://www.9news.com.au/national/alice-mccall-fashion-icon-driven-into-voluntary-administration-by-coronavirus/71f45ae8-b26d-40bb-b851-a9c33b56b08c

Norris, P. (1995). The restless searchlight: Network news framing of the post-Cold War world. *Political Communication, 12*(4), 357-370.

Orlitzky, M. (2008). The Oxford handbook of CSR.

Palmer, A. (2020a, 30 March 2020). Amazon fires warehouse worker who led Staten Island strike for more coronavirus protection. Retrieved from https://www.cnbc.com/2020/03/30/amazon-fires-staten-island-coronavirus-strike-leader-chris-smalls.html

Palmer, A. (2020b, 24 October 2020). How Amazon keeps a close eye on employee activism to head off unions. Retrieved from https://www.cnbc.com/2020/10/24/how-amazon-prevents-unions-by-surveilling-employee-activism.html

Pang, A., Jin, Y., & Cameron, G. (2010). Contingency theory of strategic conflict management: Directions for the practice of crisis communication from a decade of theory development, discovery and dialogue.

Parker, G. (2016). 5 Huge Companies Known For Implementing Horrific Working Conditions. Retrieved from https://moneyinc.com/5-huge-companies-known-implementing-horrific-working-conditions/

Pasquier, V., & Wood, A. J. (2018). The power of social media as a labour campaigning tool: lessons from OUR Walmart and the Fight for 15. *ETUI Policy Brief*.

Peterson, H. (2018, 12 September 2018). Missing wages, gruelling shifts, and bottles of urine: The disturbing accounts of Amazon drivers may reveal the true human cost of 'free' shipping in the US. Retrieved from https://www.businessinsider.com.au/amazon-delivery-drivers-reveal-claims-of-disturbing-work-conditions-2018-8?r=US&IR=T

Porter, T., & Miles, P. (2013). CSR longevity: Evidence from long-term practices in large corporations. *Corporate Reputation Review, 16*(4), 313-340.

Pound, N. (2020, 12 October 2020). What can we do about Amazon’s treatment of its workers? Retrieved from https://www.tuc.org.uk/blogs/what-can-we-do-about-amazons-treatment-its-workers

Procter & Gamble. (2020, 3 June 2020). Retrieved from https://www.instagram.com/p/CA-XFyaJBUP/

Rana, A., & Dastin, J. (2020, 30 October 2020). Amazon gets sales boost as more people shop online. Retrieved from https://www.afr.com/companies/retail/amazon-gets-sales-boost-as-more-people-shop-online-20201030-p569ys

Randewich, N. (2020, 30 April 2020). Amazon is Wall Street's biggest winner from coronavirus. Retrieved from https://www.reuters.com/article/us-amazon-stocks-idUSKBN22B2ZU

Reich, R. (2020, 13 December 2020). Jeff Bezos became even richer thanks to Covid-19. But he still won't protect Amazon workers. Retrieved from theguardian.com/commentisfree/2020/dec/12/jeff-bezos-amazon-workers-covid-19-scrooge-capitalism

Sainato, M. (2020, 7 April 2020). 'Jeff Bezos values profits above safety': Amazon workers voice pandemic concern. Retrieved from https://www.theguardian.com/technology/2020/apr/07/amazon-warehouse-workers-coronavirus-safety

Sarwar, A., & Muhammad, L. (2020). Impact of employee perceptions of mistreatment on organizational performance in the hotel industry. *International Journal of Contemporary Hospitality Management, 32*(1), 230-248. doi:10.1108/IJCHM-01-2019-0046

Scheufele, B. (2004). Framing-effects approach: A theoretical and methodological critique. *Communications, 29*(4), 401-428.

Sherman, N. (2021, 10 February). Amazon fight with workers: 'You're a cog in the system'. Retrieved from https://www.bbc.com/news/business-55927024

Simmons, L. L., Mukhopadhyay, S., Conlon, S., & Yang, J. (2011). A computer aided content analysis of online reviews. *Journal of Computer Information Systems, 52*(1), 43-55.

Sobande, F. (2019). Woke-washing:“Intersectional” femvertising and branding “woke” bravery. *European Journal of Marketing*.

Statista. (2021). Number of Amazon.com employees 2007-2020. Retrieved from https://www.statista.com/statistics/234488/number-of-amazon-employees/

Stohl, C. (1995). *Organizational communication*: Sage.

The Lancet. (2020). The plight of essential workers during the COVID-19 pandemic. *The Lancet 395*(10237), 1587.

Verrone, S. (2019, 11 January 2019). Amazon has a history of mistreating its employees. Retrieved from https://www.thetriangle.org/opinion/amazon-has-a-history-of-mistreating-its-employees/

Virgin Australia. (2020, 21 April 2020). Virgin Australia enters voluntary administration. Retrieved from https://newsroom.virginaustralia.com/release/virgin-australia-enters-voluntary-administration

Vredenburg, J., Kapitan, S., Spry, A., & Kemper, J. A. (2020). Brands Taking a Stand: Authentic Brand Activism or Woke Washing? *Journal of Public Policy & Marketing, 39*(4), 444-460.

Vredenburg, J., Spry, A., Kemper, J. A., & Kapitan, S. (2018, 5 December 2018). Woke washing: what happens when marketing communications don’t match corporate practice. Retrieved from https://theconversation.com/woke-washing-what-happens-when-marketing-communications-dont-match-corporate-practice-108035

Waddock, S. A., & Graves, S. B. (1997). The corporate social performance–financial performance link. *Strategic management journal, 18*(4), 303-319.

Wilkins, R. (2020, 26 May 2020). WHO’S HIT HARDEST BY THE COVID-19 ECONOMIC SHUTDOWN? Retrieved from https://pursuit.unimelb.edu.au/articles/who-s-hit-hardest-by-the-covid-19-economic-shutdown

Woodward, A. (2020). Amazon told workers paid sick leave law doesn't cover warehouses, report says. Retrieved from https://www.independent.co.uk/news/world/americas/amazon-paid-sick-leave-workers-warehouse-coronavirus-a9504821.html

Zamoum, K., & Gorpe, T. (2018). Crisis management: A historical and conceptual approach for a better understanding of today's crises. In Crisis Management Theory & Practice. Edited by Kattarina Holla, Michal Titko and Jozef Ristvej. London: IntechOpen Limited, pp 203-217. ISBN 978-1-78923-234-9. At: https://www.intechopen.com/books/crisis-management-theory-and-practice/crisis-management-a-historical-and-conceptual-approach-for-a-better-understanding-of-today-s-crises DOI: 10.5772/intechopen.76198. In.

MOBILE-ASSISTED SHOWROOMERS, COMPETITIVE OR LOYAL?

María Alesanco-Llorente, Aurea Subero-Navarro, Cristina Olarte-Pascual, Eva Reinares-Lara

Universidad de La Rioja (Spain), Universidad de La Rioja (Spain), Universidad de La Rioja (Spain), Universidad Rey Juan Carlos (Spain)

maria.alesanco@unirioja.es; aurea.subero@unirioja.es; cristina.olarte@unirioja.es; eva.reinares@urjc.es

INTRODUCTION

Offering a seamless shopping experience –where the barriers between online and offline world are broken (Verhoef et al., 2015)– is the maxim of the omnichannel strategy and technological devices are its great ally (Mosquera et al*.,* 2018). Devices can be part of disruptive or sustainable technology. While sustaining technology depends on the incremental improvements in the already existing technology, a technology is considered disruptive when it changes the basis of the competitive game by introducing a dimension where products did not previously compete (Danneels, 2004). In the case of commerce, disruptive technologies are constantly emerging aimed at both, supply, and demand to improve customer service and facilitate their purchasing decisions.

Smartphones are one of these disruptive technologies since many products were created solely due to its existence (Sarwar and Soomro, 2013). Furthermore, in the situation we are in, after the appearance of COVID-19, 86.2% of consumers have declared to use the smartphone inside physical shops in the different stages of the shopping journey (iVend Retail, 2019). Smartphones are used in-store to compare prices, to look for opinions about a product, to redeem discount coupons, to pay, to see how a garment fits without trying it on, etc. This behavior is known as showrooming (Sit et al*.,* 2018), where product information is collected in the physical shop and the purchase is made through online channels (Schneider and Zielke, 2020). More specifically, the smartphone-showrooming binomial leads to a new category of consumer called *mobile-assisted-showroomer (*MAS) (Sit et al*.,* 2018). It should be noted that the use of smartphones, in general, entails ethical dilemmas, such as those arising from security and data protection issues and informed consent. We must also ask ourselves about the possible change in behaviour and the autonomy of the consumer. Is the consumer more dependent on the information provided by his or her mobile phone and, therefore, less autonomous, and free? All this would influence personal identity and resource allocation.

Previous literature has distinguished between the competitive and traditional showroomer consumer (Gensler et al., 2017). The difference between them lies in the retailer where the purchase is made. While the competitive showroomer seeks product information from retailer A and buys online from retailer B, the traditional (or loyal) showroomer changes only the channel but not the retailer (Schneider and Zielke, 2020). In this regard, competitive showrooming inherently implies an ethical component (Burns et al., 2019). This behavior involves purchasing a product without a payment to the retailer from whom we take information and consumers are expected to be less likely to participate in activities they consider unethical (Babin and Babin, 1996).

In-store smartphone use is an important indicator of competitive showrooming behavior (Rapp et al., 2015). But is the new consumer MAS, competitive or loyal? Does the new consumer MAS have these ethical dilemmas? What is clear is that smartphones have modified traditional business models (Mosquera et al*.,* 2018), challenging the dominant ethical and moral patterns (Lin et al*.,* 2020).

The goal of this study is to understand MAS behavior from the under-researched ethical point of view and to answer all the questions raised. This research is based on the Multidimensional Ethics Scale (Reidenbach and Robin,1990) which considers that individuals use five dimensions to make ethical judgements: "moral equity", "relativism", "utilitarianism", "egoism" and "contractualism".

LITERATURE REVIEW

In the field of the circular evolution of ethics what an individual considers ethical influences his or her behaviour and, over time, the behaviours he or she observes influence what he or she believes to be ethical (Goel et al., 2016). What has happened in the case of the smartphone in the physical shop is that it has gone from being a socially unapproved behaviour (even in many shops there were signs prohibiting its use) to being a standardised and accepted behaviour. In a recent study we found that 79% of the Spanish population declares to use/consult the smartphone in their purchase process in the physical shop. From the point of view of applied ethics, it is known that decisions and actions are often guided by specific ethical perceptions of the context rather than an absolute consideration of what can or should be done (Cohen, 2005; LaFollette, 2002). In addition, and according to the *Psychological Contract Theory,* decision making is subjective, consumers make decisions like those made by other individuals in the absence of absolute rules of what can and cannot be done (Thompson and Hart, 2006; Goel et al., 2016).

The new omnichannel environment makes it easier and more convenient for consumers to use their smartphones at the point of sale. A decade ago, the ethical beliefs about the use of the smartphone in the shop were clearly different from those of today. Ethical beliefs depend on individual factors such as gender, age, educational level, as well as certain situation-specific factors (Ford, & Richardson, 1994).

The Multidimensional Ethics Scale (MES)

The field of ethics contains a variety of philosophies, which have their own principles for evaluating how people make ethical decisions (Kujala and Pietilainen, 2006). The Multidimensional Ethics Scale (MES) (Reidenbach and Robin, 1988, 1990) was originally derived from business ethics literature (Beauchamp and Bowie, 1983; DeGeorge,1986). Its development has allowed that the five dimensions included in the MES *"moral equity", "relativism", "utilitarianism", "egoism"* and "*contractualism*" represent the modern ethical thinking (Shawver and Sennetti, 2009). These dimensions are defined in Table 1.

The MES has been used in the context of consumer behaviour (Nguyen, 2008) and in the field of acceptance of disruptive technologies (Reinares-Lara et al., 2018; Arias-Oliva et al., 2020; Olarte-Pascual et al., 2021).

Table 1. Dimensions of the MES.

Dimensions	Concepts
Moral equity	Based on justice theory; refers to fairness, justice, rightness, and goodness (Rawls, 1971, 1999; Nguyen and Biderman, 2008; Leonard et al., 2017).
Relativism	Based on the idea that social and cultural systems help individuals define their ethical beliefs (Reidenbach and Robin, 1990). Perceptions of what is right or wrong at any level, including the organizational level, are rooted in the social and cultural system (Nguyen and Biderman, 2008; Duran and Park, 2014).

Egoism	Based on teleological or consequentialist theories, in which morality is measured by the consequences of actions on individuals (Reidenbach and Robin, 1990). Analyses long-term individual self-interest, including aspects related to self-promotion and personal satisfaction (Nguyen and Biderman, 2008; Park, 2014; Leonard et al., 2017).
Utilitarianism	Based on teleological or consequentialist theories, in which morality is measured by the consequences of actions on society (Reidenbach and Robin, 1990). Analyses the aggregate social cost and benefits (Nguyen and Biderman, 2008; Berger et al., 2008; Park, 2014).
Contractualism	Based on a purely deontological dimension including notions of implied obligation, contracts, duties, and rules (Reidenbach and Robin, 1990). Perceptions of what is right or wrong are based on notions of an implied non-written or non-spoken contract that exists between business and society and influences all behavior (Nguyen and Biderman, 2008; Leonard et al., 2017).

Source: Arias-Oliva et al*.,* 2020

METHOD

The information was collected through semi-structured personal surveys applied to a sample of 217 Spanish MAS consumers who use their smartphones at physical clothing stores to look for information (64.1%), compare prices (52.1%), compare products (51.2%), read reviews by other shoppers (31.7%), share photos (68.7%), redeem coupons (65.9%), or pay (44.7%). Additionally, a quantitative and qualitative information treatment was carried out. The characteristics of the sample are: Men 32.7% and Women 67.3%; Ages between 18-25 years old, 13.8%, 26-35 years old, 21.2%, 36-45 years old, 21.7%, 46-55 years old, 23.9% and over 56 years old, 19.4%.

The scale used to measure ethical judgements have been adapted from Composite MES by Shawver and Sennetti (2009). All items were measured using a Likert scale ranging from 0 to 10. For example, Unjust-Just: 0 refers to Unjust; 5 is the midpoint between Unjust-Just and as it approaches 10 it is because it is considered Just. Likewise, an open question is posed to know the attitude of the participants to the study scenario.

It is important to pay attention to the moral scenario the respondents are asked to assess under the MES scale items or moral principles and an open question (Table 2). The scenario was developed based on the behavior with the highest representation during the purchase process (Schneider and Zielke, 2020) and more controversial from the ethical point of view: competitive showrooming.

This scenario describes a situation where a person is in a store trying on a shirt that she/he likes and checks her/his smartphone to see if it is cheaper online. Once she/he finds out that is cheaper online in another store, she/he decides to buy it in the other online store with his smartphone and leave the physical store without buying.

Table 2. The MES scale ítems.

Moral Equity	**EQ**
Unjust/Just	EQ1
Unfair/Fair	EQ2
Not morally right/Morally right	EQ3
Relativism	**R**
Not acceptable to my family/Acceptable to my family	R1
Culturally unacceptable/Culturally acceptable	R2
Traditionally unacceptable/Traditionally acceptable	R3
Egoism (E)	**E**
Not self-promoting for me/Self-promoting for me	E1

Utilitarianism (U)	**U**
Produces the least utility/Produces the greatest utility	U1
Minimise benefits and maximise hurt/ Maximise benefits and minimise hurt	U2
Contractualism (C)	**C**
Violates/does not violate an unwritten contract	C1
Violates/does not violate an unspoken promise	C2

RESULTS

When participants were asked whether they would act in the same way, results show that there is a clear division in the responses to the scenario. The results obtained according to the age and gender of the participants justifies the ethical controversy produced by this type of behavior (Table 3). Both for men and women it is observed that the higher the age range, the lower the affinity with the described behavior.

Participants (P) who support this behavior put their economics before their own ethical judgments. *It is the Law of supply and demand. I worry about my finances* (P19). *For me there is no ethical problem, in these times everyone looks out for their own economic interest* (P68). Meanwhile, 32.3% of participants did not observe any ethical problems. *In my opinion there is no ethical problem, nowadays technology and the internet are used to make a profit, in this case if you can get the clothes you like for a lower price, I do not think it will hurt anyone* (P206).

For their part, the participants who do not support it say that this type of behavior destroys small local businesses and reduces the number of jobs in the commercial sector and that it is also a lack of respect for the seller of the physical store. *That street commerce disappears and with it the center of our cities and many jobs* (P140). *Disrespect towards workers, they try to help you and you leave them in the background* (P211). *Selling online does not employ people and can cause problems for the structure of our cities* (P103).

Table 3. Distribution of agreeing and disagreeing in the scenario.

Scenario	**Male**		**Female**	
	Agreeing	**Disagreeing**	**Agreeing**	**Disagreeing**
18-25	75.00%	25.00%	61.54%	38.46%
26-35	56.25%	43.75%	50.00%	50.00%
36-45	42.86%	57.14%	57.69%	42.31%
46-55	35.71%	64.29%	34.21%	65.79%
> 56	18.75%	81.25%	23.08%	76.29%

Then the Tables means are shown for the items of the MES scale grouped by age and sex. Results for men are highlighted in blue and for women, in green. The vertical red line represents the global mean of the item.

The Results for the "*Moral Equity*" dimension (Tables 4-6) reveal that men consider competitive showrooming behavior fairer, more just, and more morally correct than women for the first four age ranges (18-55 years). However, the trend changes for the last range, where the average of women exceeds that of men.

Table 4. Means for EQ1 by age and gender.

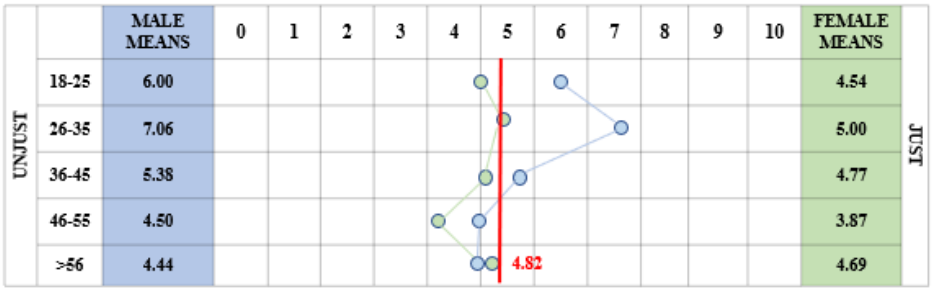


Table 5. Means for EQ2 by age and gender.

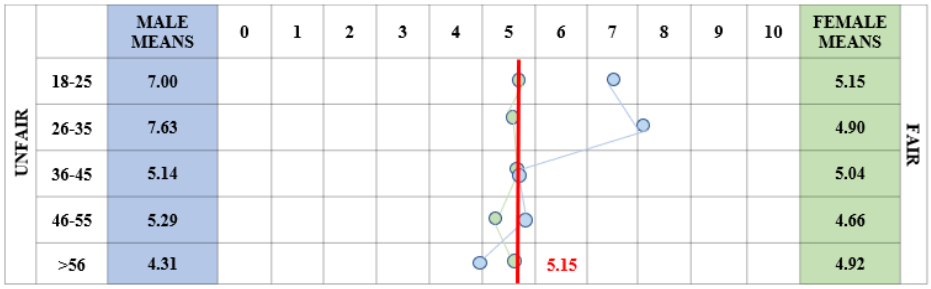
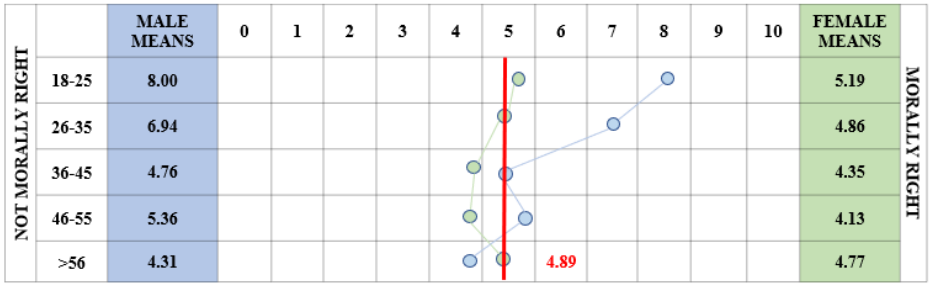


Table 6. Means for EQ3 by age and gender.



For all items of "*Relativism*", the pattern drawn by both genders in the three highest age ranges is similar, however, men show a slightly higher mean. The most outstanding thing is to observe how men between 26 and 35 years old have very high means compared to the rest of the groups and the global mean (Table 7-9). Therefore, their perception of what is cultural, traditional, and familiarly accepted doesn´t help competitive showrooming to be valued in a more ethical way when making an ethical judgment.

Table 7. Means for R1 by age and gender.

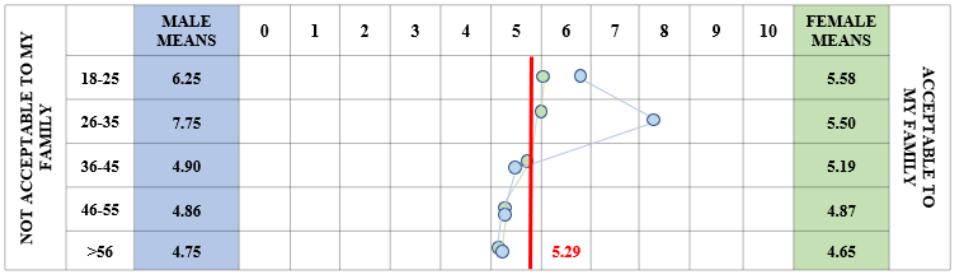


Table 8. Means for R2 by age and gender.

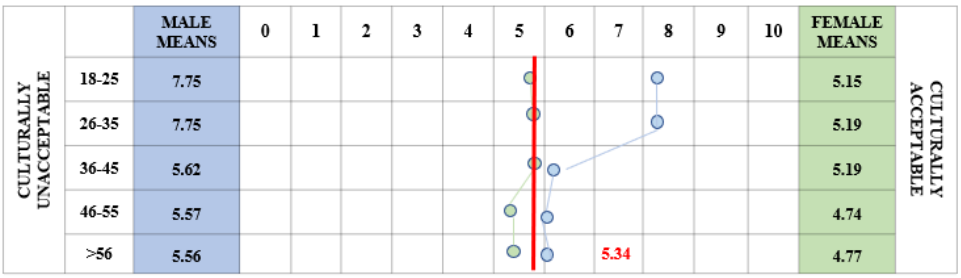
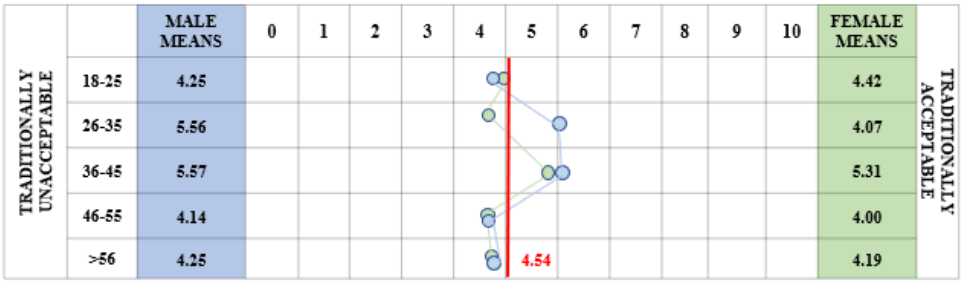
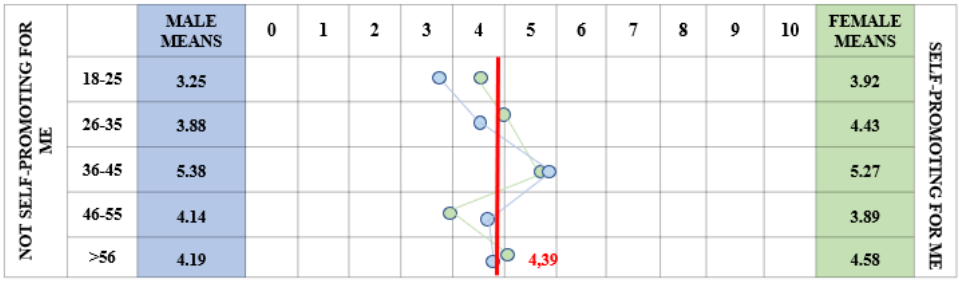


Table 9. Means for R3 by age and gender.



The means of the item that makes up the dimension "*Egoism*" present different results to the patterns presented in the previous Tables (Table 10). For the first time, women show upper means in the first two age ranges. In this sense, we could say that young women measure less the individual consequences of this type of buying behavior.

Table 10. Means for E1 by age and gender.



The means for "*Utilitarianism*" dimension reappeared in the group of men between 26 and 35 years old (8.38 for U1) (Table 11-12). The higher the mean, the lower is how they consider the consequences of this behavior in society to form the total morality of the act.

Table 11. Means for U1 by age and gender.

PRODUCES THE LEAST UTILITY		MALE MEANS	0	1	2	3	4	5	6	7	8	9	10	FEMALE MEANS	NOT PRODUCES THE GREATEST UTILITY
	18-25	5.00												6.58	
	26-35	8.38												6.00	
	36-45	6.10												5.46	
	46-55	6.43												5.92	
	>56	5.69							6.07					5.27	

Table 12. Means for U2 by age and gender.

MINIMISE BENEFITS AND MAXIMISE HURT		MALE MEANS	0	1	2	3	4	5	6	7	8	9	10	FEMALE MEANS	MAXIMISE BENEFITS AND MINIMISE HURT
	18-25	5.00												5.35	
	26-35	6.06												5.33	
	36-45	5.14												5.04	
	46-55	4.64												4.97	
	>56	5.44							5.10					4.23	

The perceptions of what is right or wrong measured by the means of the "*Contractualism*" show: (1) for C1 (Table 13), the youngest men group doesn´t have a strong notion about the relation between business and society (8.25), (2) for C2 (Table 14), for the first time, women with higher age ranges present higher means than those presented by young women.

Table 13. Means for C1 by age and gender.

VIOLATE AN UNWRITTEN CONTRACT		MALE MEANS	0	1	2	3	4	5	6	7	8	9	10	FEMALE MEANS	DOES NOT VIOLATE AN UNWRITTEN CONTRACT
	18-25	8.25												4.81	
	26-35	5.19												6.20	
	36-45	5.81												5.58	
	46-55	6.07												4.55	
	>56	5.81						5.28						3.88	

Table 14. Means for C2 by age and gender

VIOLATE AN UNSPOKEN PROMISE		MALE MEANS	0	1	2	3	4	5	6	7	8	9	10	FEMALE MEANS	DOES NOT VIOLATE AN UNSPOKEN PROMISE
	18-25	6.00												4.81	
	26-35	4.69												5.67	
	36-45	6.14												5.42	
	46-55	6.00												4.82	
	>56	5.75								5.24				4.35	

CONCLUSIONS

The purpose of this study was to measure the ethical judgments of the new consumers mobile-assisted showroomers. The most relevant contributions are related to the questions that were raised at the beginning of the study:

- Is the new consumer MAS, competitive or loyal?

  The results of their perceptions against the scenario proposed in the study –where a competitive showrooming behavior is observed– allows us to understand if we are facing a new competitive trend (looking for product information from retailer A and buy online from retailer B (Schneider and Zielke, 2020)) or a loyal one (changing the purchase channel but not the retailer (Schneider and Zielke, 2020)).

  Although it is impossible to generalize a behavior to the entire population (Giménez, 2012), it is possible to establish a pattern through the degree of agreement (or disagreement) of the participants. There seems to be a clear relationship between age and the tendency towards being and/or approving of competitive showrooming behavior. The younger the age, the greater the approval of this purchase behavior. For those who approve this conduct, the ethical image of the act itself is relegated to the background, only valued after the economic interest of the person.

- Does the new consumer MAS have these ethical dilemmas?

  All dimensions of the MES scale affect the ethical judgments of the MAS, to a greater or lesser extent. The global means for each item are close to the intermediate point of the scale (5), which indicates a clear indecision when assessing a certain character of the scale (Ethically indifferent). "*Utilitarianism*" is the most important dimension for MAS consumers. One possible explication to this is that when more is known about smartphones, ethical judgments focus more on whether they are useful in terms of their benefits vs their associated costs and inconvenience.

  However, ethical judgments are clearly conditioned by the age and gender of the sample. In general, for both genders, from the age of 36 they consider competitive showrooming less ethical than the lower age groups (Ethically against). It has been observed that ethical judgment is directly proportional to age. The group of men from 18 to 35 years old appears as the most related to competitive behavior (Ethically in favor**)**, as they show high averages compared to the rest of the group.

  Regarding the limitations of the study, the one derived from the variables included in the MES scale should be noted, since there are other factors that may condition the ethical judgments that have not been considered in this research. On the other hand, it might have been

convenient to ask respondents to tell us two shopping experiences, one positive and the other negative. In this way, we would have avoided the possible bias in the ethical judgments that occurs when the purchasing situation analyzed corresponds to a specific scenario. Future research lines could include these two suggestions.

KEYWORDS: Mobile-Assisted-Showroomer, Showrooming, Multidimensional Ethics Scale, Smartphone.

REFERENCES

Arias-Oliva, M., Pelegrin-Borondo, J., Lara-Palma, A. M., & Juaneda-Ayensa, E. (2020). Emerging cyborg products: An ethical market approach for market segmentation. *Journal of Retailing and Consumer Services*, *55*, 102140.

Burns, D. J., Gupta, P. B., & Hutchins, J. (2019). *Showrooming: the effect of gender. Journal of Global Scholars of Marketing Science, 29(1), 99 113.* https://doi.org/10.1080/21639159.2018.1551725

Cohen, A.I. (2005). Contemporary debates in applied ethics. Wiley-Blackwell.

Danneels, E. (2004). Disruptive technology reconsidered: a critique and research agenda. J. Prod. *Innovation. Management,* 21 (4), 246–258. https://doi.org/10.1111/j.0737- 6782.2004.00076.x

Ford, R. C., & Richardson, W. D. (1994). Ethical decision making: A review of the empirical literature. *Journal of business ethics*, *13*(3), 205-221.

Gensler, S., Neslin, S.A., & Verhoef. P.C. (2017), The showrooming phenomenon: It´s more than just about price. *Journal of Interactive Marketing*, Vol. 38, pp. 29–43. https://doi.org/10.1016/j.intmar.2017.01.003

Goel, L., Hart, D., Junglas, I., & Ives, B. (2016). Acceptable IS Use: Conceptualization and measurement. Computers in Human Behavior, 55, 322-328. https://doi.org/10.1016/j.chb.2015.09.029

iVend Retail, 2019, "Global Shopper Trends Report", disponible en: https://support.ivend.com/userportal?id=doc_view&doc=KB0011245 (accessed January 10, 2021).

Lazaris, C., Vrechopoulos, A. P., Doukidis, G. I., & Fraidaki, A. (2015), "Mobile apps for omnichannel retailing: revealing the emerging showrooming phenomenon", in 9th Mediterranean Conference on Information Systems (MCIS), Samos, Greece, p. 12.

Levav, J., & Mcgraw, A. P. (2009), Emotional accounting: How feelings about money influence consumer choice. Journal of Marketing Research, Vol. 46 No: 1, pp. 66-80. https://doi.org/10.1509/jmkr.46.1.66

Lin, W.L., Yip, N., Ho, J.A. & Sambasivan, M. (2020). The adoption of technological innovations in a B2B context and its impact on firm performance: an ethical leadership perspective. *Market. Management*. https://doi.org/10.1016/j.

Mosquera, A., Juaneda-Ayensa, E., Olarte-Pascual, C., & Pelegrín-Borondo, J. (2018), Key Factors for In-Store *Smartphone* Use in an Omnichannel Experience: Millennials vs. Nonmillennials, *Complexity*, Vol. 2018, Article ID 1057356 https://doi.org/10.1155/2018/1057356

Nguyen, N.T., Basuray, M.T., Smith, W.P., Kopka, D. & McCulloh, D. (2008) Moral issues and gender differences in ethical judgment using Reidenbach and Robin's multidimensional ethics scale:

implications in teaching of business ethics. J. Bus. Ethics 77 (4), 417–430. https://doi.org/10.1080/10508422.2012.672907

Olarte-Pascual C., Pelegrín-Borondo J., Reinares-Lara E. & Arias-Oliva M. (2021). From wearable to insideable: Is ethical judgment key to the acceptance of human capacity-enhancing intelligent technologies? *Computers in Human Behavior*, 114, 1-11. https://doi.org/10.1016/j.chb.2020.106559.

Reidenbach, R. E., & Robin, D. P. (1990). Toward the development of a multidimensional scale for improving evaluations of business ethics*. Journal of Business Ethics*, 9(8), 639-653. https://doi.org/10.1007/BF00383391

Reinares-Lara, E., Olarte-Pascual, C., & Pelegrín-Borondo, J. (2018). Do you want to be a cyborg? The moderating effect of ethics on neural implant acceptance. *Computers in Human Behavior*, *85*, 43-53.

Schneider, P. J., & Zielke, S. (2020), Searching offline and buying online–An analysis of showrooming forms and segments, *Journal of Retailing and Consumer Services*, Vol. 52, Article ID: 101919. https://doi.org/10.1016/j.jretconser.2019.101919

Sit, J. K., Hoang, A., & Inversini, A. (2018), Showrooming and retail opportunities: A qualitative investigation via a consumer-experience lens, *Journal of Retailing and Consumer Services*, Fernández Vol. 40, pp. 163-174. https://doi.org/10.1016/j.jretconser.2017.10.004

Thompson, J., & Hart, D. (2006). Psychological contracts: a nano-level perspective on social contract theory. Journal of Business Ethics, 68(3), 229-241. https://doi.org/10.1007/s10551-006-9012-x

ETHICS, MARKETING AND TECHNOLOGY: A CASE STUDY IN HIGHER EDUCATION IN SPAIN

Mario Arias-Oliva, Teresa Pintado Blanco, Antonio Pérez-Portabella, Araceli Rodríguez Merayo

Complutense University of Madrid (Spain), Complutense University of Madrid (Spain), Universitat Rovira i Virgili (Spain), Universitat Rovira i Virgili (Spain)

mario.arias@ucm.es; tpintado@ucm.es; antonio.perezportabella@urv.cat; araceli.rodriguez@urv.cat

ABSTRACT

Digital marketing is an emerging discipline. The application of all technological disruptions in the marketing field is transforming both the research and professional arenas. Technology provokes a revolution in traditional marketing strategies and techniques, arising many ethical concerns. Our research question is the following: are business schools teaching the ethical implications of digital marketing? The first section of the it is introduced with some examples the importance of the ethical considerations in digital marketing. It continues with the proposed methodology to analyse the ethical digital marketing competencies that future professionals are acquiring in Spain. Our findings show that there is an important gap in marketing Master's degrees with regard to the ethical aspects of digital marketing.

INTRODUCTION: WHY IS IMPORTANT ETHICAL ISSUES IN DIGITAL MARKETING EDUCATION?

The American Marketing Association (AMA, 2017) defines marketing as "the activity, set of institutions, and processes for creating, communicating, delivering, and exchanging offerings that have value for customers, clients, partners, and society at large", focusing on obtaining value for organizations' stakeholders; likewise, in the "AMA Statement of Ethics" (AMA, n.d.) it promotes the highest standard of professional ethical norms and values, focused on what is desirable, important and morally proper. All this reveals that ethics has become of great relevance to the marketing discipline.

The technological impact of Information and Communication Technologies on marketing over the last decades is enormous. Our paper focuses on the analysis of higher education in Spain, specifically about how ethical issues in digital marketing are integrated in curricula and competences.

In this section we will introduce the importance of the ethical aspects in digital marketing. It would be difficult and not relevant to our educational purpose to list the specific areas of ethical impact when using new digital tools in marketing, as this paper does not intend to create an ethical taxonomy of digital marketing. Due to this constraint, we ground the relevance of our paper with some examples that point out the importance of ethics in digital marketing, justifying why ethics should be included in the curricula of higher education for future marketing professionals.

Impact of digital tools on marketing can be classified in two broad categories:

- Impact on traditional marketing tools and strategies such as segmentation, positioning, pricing, product, promotion, etc.
- Impact on new marketing tools and strategies such as SEO, SEM, Big Data, etc.

Pricing strategies are an excellent example of technological disruption in marketing. Pricing is analysed and taught in any marketing course all over the word. All marketing handbooks incorporate a pricing chapter (e.g. Kotler & Amstrong, 2018; Kerin & Hartley, 2019; Marshall & Johnston,2015) or specific books and papers can be found among the recommend references (e.g. Schindler, 2012; Baker Benmark, Chopra, Kohli, 2018; Simchi-Levi, 2017; Liozu, 2019). But technology is transforming the pricing strategies. Marketing pricing strategies are moving from a static way of fixing prices to dynamic pricing, where prices can be now personalized for specific segments, microsegments or even for each specific customer. The price discrimination was detected in online environments in 2010. Amazon was showing different prices for the same product. In 2012, the Wall Street Journal found that prices were different based on the geographic location of customers (Mattioli, 2012). From this discrimination in prices, the new technologies have opened new amazing possibilities for dynamic pricing. The creation of dynamic prices is done by an algorithm that can change price every second depending on the location, demand, time or any other criteria defined in the algorithm. These techniques were born in one of the first electronic markets, the GDS (Global Distribution Systems) in the air travel industry (Schmid, 1994). Since then, it has been applied in many other sectors such as the energy sector (Goutam, & Krishnendranath, 2017), online advertisement auctions (Google, 2020) or public transport services such as Uber (Martin, 2019).

In the previous examples, the combination of traditional pricing policies with the possibilities of new technologies opens important opportunities in marketing, but some problems and ethical concerns arise as well. According to Veeraraghavan (2016), the use of dynamic pricing could anger frequent customers, or provoke a shift in demand towards the very last-minute bargain. In a piece of research about the effects of a dynamic pricing strategy for a concert ticket, it was shown that a significant last-minute discount creates controversy: some people find it favourable, many others hate it, and some of the customers consider it as a right practice. In the case of Uber, the price is changing constantly looking for the maximum price that the consumer is willing to pay. The algorithm establishes a price based on where you are, what time it is, how many people are demanding the service, the day of the week and the month, and also on your historical records using the Uber app (Martin, 2019). But Uber goes further and knows that on a rainy day a consumer is more likely to accept a higher price than on a sunny day, or that a key variable to accept a higher price is the level of battery of your Smartphone (Kosoff, 2016). When the battery is very low, the probability of accepting a higher price is very high because the user can miss the connectivity of their device. Using all these digital marketing strategies, companies have the ability to collect, keep and use information on consumer behaviours not in favour of their customers, but in their own interest (Calo & Rosenblat, 2017).

The price can be different depending on the operating system of your device. According to Kingsley-Hughes (2012), the online travel agency Orbiz showed different hotels and different prices to Mac users than to other users with other operative systems (e.g. Linux or Windows). Based on their records, they knew that Mac users spent 30 percent more per night for a hotel room than Windows users. Offers and prices were different for customers accessing from Mac or other devices. These techniques are known as price steering and price discrimination (Hannak, Soeller, Lazer, Mislove & Wilson, 2014). Price steering occurs when different products are shown to the customer (or the order of showing products is different for each customer) depending on their profile, and it becomes a usual strategy in online environments. Pemberton, Stonehouse & Barber (2001) pointed out their concerns in the air travel industry. GDS creates an “halo effect” which makes customers usually choose products that appear in the first position of the screen. A consumer usually does not ask for a hundred flights before choosing the best option in a flight, he/she asks for a few combinations and decides. That “halo effect” allows companies to control the system and manipulate it, showing on the first screen the options of the airlines that own the GDS. Companies alter competition with unethical behaviour.

Business has all the information about both individual customer preferences and aggregated customer preferences. But customers know nothing about the information that the companies have and use in their own interest. For instance, whether the company must sell products fast because a new version is going to be launched soon. This asymmetry in the information creates important differences.

The need to manage data ethically was discussed as early as the 1990s. The 21st century illuminated the way to multiple ethical codes: Institute of Electrical and Electronics and Association for Computing Machinery are the most important. But self-regulation is not enough. Couldry & Turow (2014) conclude that incorporating big data into personalized marketing and content production threatens the ecology of social relationships. Therefore, the misuse of mass information could be a democratic threat. There are many economic incentives in controlling big data. The companies at the top positions in global stock market valuations 2020 are technology intensive: Microsoft, Apple, Amazon, Amazon, Alphabet, Alibaba, Facebook, or Tencent.

The Facebook and Cambridge Analytica gate demonstrated the need to regulate the use of private data in Europe. But the lobby CCIA Europe (@CCIAEurope) advocate for a free and open Internet in Europe and is against the European regulation and tries to exert influence on the European institutions as Transparency International reports.

Data is an essential part of technology-intensive business models. The dynamics of data flows can shape political communities, influence democratic elections, and build or destroy reputations, individual and group self-concepts (Islam, 2021). Price discrimination and health prediction (e.g., insurance policies) could become ethically problematic if people are denied access to essential goods or services based on their income or lifestyle (Favaretto et al., 2019). Experts point out that incorporating big data into personalized marketing threatens the ecology of connections that bind citizens and groups because it influences their information and empathy (Couldry & Turow, 2014). Studies shows the communicative power of new media in promoting ethical causes (Banaji & Buckingham, 2009).

The Internet of Things brings countless benefits to people's well-being, health care and productive advances in industry, but the Artificial Intelligence (AI) must also serve democracy and human rights but will not serve the public good without strong rules in place (Nemitz, 2018).

The media influences social norms, consumer choices, or ethical consumption. The debate is whether social media should follow the market trend ethically, which does; or it must exempt that ethics from the social patterns on which it exerts influence (Lekakis, 2014). The use of Big Data must be carried out under ethical criteria: it is necessary to assess the profound impact that actions arising from its observation have on the object analysed itself (Boyd & Crawford, 2012)

The accumulation of digital power, which shapes the development and deployment of AI as well as the debate on its regulation, is based on four sources of power (Nemitz, 2018): (1) money being the classic tool of influence on politics and markets; (2) these corporations increasingly control the infrastructures of public discourse and the digital environment decisive for elections; (3) the mega corporations know more about us than ourselves; (4) these corporations are dominating development and systems integration into usable Artificial Intelligence services.

Ethical implications arise from the use of personal data to make decisions—whether policies, planning, or resource allocation—that affect entire populations based on the data of a few (Crawford et al., 2014). Crawford et al. describes 'marketing comfort' as the link between marketing ethics and consumer confort (2020:7).

Consumers perceive information collected about themselves such as exchange of online services, goods, or something more of value (Ashworth & Free, 2006). The risks and benefits perceived by

consumers from using social media relate to their convenience to sellers who use their publicly available social media data (Jacobson et al., 2020).

Some of the conditions of trust, especially security and privacy, are fundamental human needs and are preconditions for the development of autonomous moral humans (Ahearne et al., 2005).

The intense relationship between e-commerce and trust has been profusely studied (Maximilien & Singh, 2005; Tomlinson & Mayer, 2009). Trust is the basis to forging and maintain long-term e-commerce relationships (Sharma & Lijuan, 2014), and as increase transaction complexity makes conditions more uncertain –as is in computer-mediated business– the need of trust increases. E-commerce ethics was a direct influencing factor of trust, security and privacy and loyalty (Sharma & Lijuan, 2014).

TRAINING THE FUTURE DIGITAL MARKETING PROFESSIONALS IN ETHICS

Learning responsibility to the community is more than just mere voluntarism, it is necessary. This formative intention should be incorporated into the academic curriculum corpus and university culture (Buxarrais & Esteban, 2004). In the age of technique learning profession is in danger of not reaching the necessary depth in their professional scope. The university student can easily become a piece that has lost the meaning of his work, can become one more terminal of the production system (Castells, 1998). University institutions are committed to the construction of the professional in all its fields: techniques, attitudes, and responsibilities. Comprehensive training is increasingly needed to enable teachers to face an uncertain and complex society and the new profiles of students arriving at the university (Montes & Suárez, 2016). The digital necessity in marketing education has increased in four new areas chronically under-taught in universities (Crittenden & Crittenden, 2015; Moscoso Pozo et al., 2017): search engine optimization, social media, marketing software skills, and online-lead generations strategies.

Fourali (2009) described the process of developing best practice standards for social marketing in the United Kingdom by the Marketing and Sales Standards Setting Board (MIC): Used existing standards and identify the specific needs there may be for social marketers.

The integration of ethical training in the university requires a change in the teaching culture of teachers: ethical training should be provided to future professionals to know the duties and obligations in the practice of the profession (González Pérez et al., 2014). If a professional is legitimized as an expert, professional competence is not sufficient, it is necessary to make the commitments he/she shares with his/her colleagues (Bolivar, 2005).

In recent years, Corporate Social Responsibility (CSR) has been incorporated into university curriculums (Manuel Larrán et al., 2014). The new Marketing DNA (Harrigan & Hulbert, 2011) incorporates the new digital reality of communication and commerce. The corporate digital responsibility (CDR) culture relates to digital responsibility and embodies shared values from which specific CDR standards are derived that then lead to specific behaviours (Lobschat et al., 2021).

Based on previous findings and examples, we question how ethical those new marketing tools and techniques are? Does a business using all these techniques inform their consumers? To what extent is ethical to control the market in the company's own interest? Making aware of ethical aspects to future digital marketing professionals is a must. This research explores how ethical aspects of digital marketing are integrated into marketing higher education.

METHOD

Sample selection

A wide variety of university master's degrees are currently being offered in Spain in marketing that include specific content in this subject, or that are shared with other related areas. According to the data provided by the Spanish Ministry of Science, Innovation and Universities (20 April 2021), 157 Master's degrees are offered based on searches for the following keywords in their titles: "marketing", "communication", "advertising", "consumer", "commercial" and "trade", according to the distribution presented in Table 1; however, the terms "communication", "commercial" and "trade" in other areas that have not been taken into account (e.g., Master's degrees in telecommunications). Likewise, the terms "market", "marketing" and "consumer" were searched, but the results were already included in the previous searches, while the terms "distribution" and "marketing" did not produce any results.

Table 1. University master's degrees offered in the area of marketing.

Key word in the title of the master's degree	No. of university master's degrees	Type of university (public/private)	Online programs	Taught in a foreign language
Marketing	63	31 public / 32 private	18	12
Communication	70	38 public / 32 private	23	4
Advertising	8	3 public / 5 private	3	0
Consumer	4	2 public / 2 private	0	2
Commercial	7	1 public / 6 private	4	1
Trade	5	2 public / 3 private	1	0
TOTAL	**157**	**77 public / 80 private**	**49**	**19**

Source: self elaboration based on Spanish Ministry of Science, Innovation and Universities (20 April 2021. https://www.educacion.gob.es/notasdecorte/busquedaSimple.action)

As can be seen in Table 1, the Master's degrees are similarly distributed between public and private universities. Most of the 157 Master's degrees face-to-face format, with only 49 degrees offered exclusively online. Considering that the methodological approach of the research is qualitative, we select for the exploratory research five Masters that focus specifically in digital marketing. The selected sample can be seen in Table 2.

Table 2. Sample selected.

Master	University
Digital Marketing Management	University of Málaga (UMA)
Digital Marketing	University of Mondragon (UMO)
Digital Marketing Analysis	University of Murcia (UM)
Digital Marketing	Open University of Catalonia (UOC)
Social Media and Strategic Management	Open University of Catalonia (UOC)

Source: self-elaboration.

For each of the selected sample Masters, we analize the Official Guide that Master used to approve the program in the Official Spanish Government Office (ANECA, Agencia Nacional de Evaluación de la Calidad y Acreditación). A Content Analysis method was used. Content analysis “is a research tool used to determine the presence of certain words, themes, or concepts within some given qualitative data (i.e. text). Using content analysis, researchers can quantify and analyze the presence, meanings and

relationships of such certain words, themes, or concepts" (Columbia Public Health, 2021). In content analysis we focus on competencies. The OCDE in its Definition and Selection of competences (OCDE, 2002: 4) defines a competence as "the ability to respond to demands or carry out tasks successfully and consistently with cognitive and non-cognitive dimensions". Within this general framework, we can define transversal competences as those skills related to personal development, which appear in all domains of professional and academic performance (González and Wagenaar, 2003). It is a very complex know-how, which is why it is necessary to specify more specific learning outcomes. In contrast to these competences, we find the specific competences that refer to those skills that are necessary for the performance of the work in question. Within the transversal competences, which are those offered by the Spanish Universities to all degrees, we find the basic competences that are common to all degrees at the same MECES level and are established in section 3.3 (Master's degree) by Royal Decree 861/2010, of 2 July, and in article 5 of Royal Decree 99/2011. A competence is therefore a set of learning outcomes. A learning outcome is a written statement of what the learner is expected to be able to do at the end of a module, subject or course.

Our research focuses on Official Guides of each Master program, focusing on how ethics is integrated into competences definitions. We analyzed as well if ethical aspects are in other parts of the Master program such as subjects, learning outcomes, references or other descriptions.

Results

In this study, we have analysed a sample of 5 university Master's degrees in digital marketing in order to check the mentions of the word "ethics" or "ethical" in the different contents of their Verification Reports (guides presented for program approval to Spanish Government responsible agency). As can be seen in Table 3, there are a few contents related to these terms: the largest number of mentions is found in the learning outcomes (4 mentions over 3 Master's degrees) and in the publications included in the report (4 mentions, although 3 refer to the same publication). The remaining mentions are found in a justification, in subjects and their descriptors, and in specific headings or in other sections of lesser relevance.

Tabla 3. Frequencies of terms related to ethic.

	Subject	**Subject heading**	**Subject description**	**Learning outcome**	**Publication**	**Descriptions**	**Others**
Digital Marketing Management (UMA)	0	0	0	1	0	0	0
Digital Marketing (UMO)	1	2	0	0	1	0	3
Digital Marketing Analysis (UM)	1	1	2	0	0	0	1
Digital Marketing (UOC)	0	0	0	1	0	0	0
Social Media and Strategic Management (UOC)	0	1	0	2	3	1	2

Source: self-elaboration based on content analysis results

However, the importance of ethics in the Master's degrees can be seen mainly in the mentions that ethical related terms have in the competences. Table 4 shows that the frequencies in which the term "ethics" or "ethical" appears on competences of selected Masters. It is included 10 occasions in all the degrees analyzed, always as a basic competence, and occasionally as a general, transversal or specific competence.

Table 4. Frequency of ethic related terms in the competencies of the Verification Reports.

Number of competencies					
Masters	Basic	General	Transversal	Specific	Total
Digital Marketing Management (UMA)	1	1	0	1	3
Digital Marketing (UMO)	1	0	0	0	1
Digital Marketing Analysis (UM)	1	0	0	1	2
Digital Marketing (UOC)	1	0	1	0	2
Social Media and Strategic Management (UOC)	1	0	0	1	2

Source: self-elaboration based on content analysis results

But these competences mentioned are repeated on multiple occasions throughout the Verification Reports, as shown in Table 5. Thus, in the first degree analysed (UMA) it can be seen that a total of 3 different competences appear (Table 4), although these are repeated a total of 41 times (Table 5), therefore, the competences that include ethic related terms are not varied, although they are very often repeated.

Table 5. Frequency of ethic related terms in competencies.

Competencies					
Masters	Basic	General	Transversal	Specific	Total
Digital Marketing Management (UMA)	25	15	1	0	41
Digital Marketing (UMO)	1	0	0	0	1
Digital Marketing Analysis (UM)	4	0	o	5	9
Digital Marketing (UOC)	5	-	7	-	12
Social Media and Strategic Management (UOC)	3	0	0	5	8

Source: self-elaboration based on content analysis results

CONCLUSIONS

The results of our study reveal that in the sample of Master's degrees analysed there is no a generalized integration of ethics in digital marketing studies in Spain, answering our research question (are business schools teaching the ethical implications of digital marketing?) we find that ethics is only incorporated occasionally in the analysed cases. This finding contrasts with the literature reviewed, which mentioned the intention to include academic content on responsibility into the academic curriculum corpus and university culture (Buxarrais & Esteban, 2004) and the need to train professionals prepared to apply the corporate digital responsibility (CDR) (Lobschat et al., 2021). However, ethics appears timidly in the degree competencies, and it is foreseeable that in the future its use will be of greater interest to higher education institutions and will be transferred to a greater number of degree competencies.

The main limitation of our study is the use of a limited sample of Master's degrees, although the aim was not to carry out an extensive work, but to an exploratory study in the field, doing as a very first approach to know the situation of ethics in Master's degrees in Spain. On the other hand, our work has focused on the analysis of the content analysis of Verified Reports that serve for the implementation of the degrees, although it would be interesting to study other complementary materials and methods to verify the application of ethics in the usual teaching, although in this case it would be necessary to have the support of the university institutions to obtain more information. Likewise, the ethical training in the university requires a change in the teaching culture of teachers (Gozálvez Pérez et al., 2014), although in this first phase of our study we have not contacted them, contemplating it as a later objective.

The future lines of our work can be oriented to the analysis of a larger number of Masters, to expand the sample and obtain more conclusive results. Likewise, the analysis could include other educational levels that teach digital content in the field of marketing (Degrees, Vocational Training, among others), and even carry out a comparative study between them. Finally, a line of great interest would be the analysis of specific information from educational institutions, as well as a study among teachers to analyse how they apply ethics in the delivery of their teaching and how it is transferred to the learning outcomes of students.

ACKNOWLEDGEMENTS

Work produced within the FLOASS project - Learning Outcomes and Learning Analytics in Higher Education: An Action Framework from Sustainable Assessment (Resultados y analíticas de aprendizaje en la educación superior: un marco de acción desde la evaluación sostenible), funded by the Ministry of Science, Innovation and Universities in the Spanish R+D+i Programme Focussed on Challenges to Society and the European Regiona lDevelopment Fund (Ref. RTI2018-093630 -B -100) and by the innovation project ACCRAM – Análisis de la Calidad de las Competencias y Resultados de Aprendizaje de los Másteres (Ref. INDOC19-07GI1926), funded by the Call for Teaching Innovation Projects of the Rovira i Virgili University (URV).

KEYWORDS: digital marketing, ethical marketing, higher education, ethical competences.

REFERENCES

Ahearne, M., Bhattacharya, C. B., & Gruen, T. (2005). Antecedents and consequences of customer-company identification: Expanding the role of relationship marketing. *Journal of Applied Psychology*, 90(3), 574–585. https://doi.org/10.1037/0021-9010.90.3.574

American Marketing Association – AMA (2017). *Definitions of marketing.* https://www.ama.org/the-definition-of-marketing-what-is-marketing/

American Marketing Association – AMA (n.d.). *Codes of Conduct | AMA Statement of Ethics.* https://www.ama.org/codes-of-conduct/

Ashworth, L., & Free, C. (2006). Marketing dataveillance and digital privacy: Using theories of justice to understand consumers' online privacy concerns. *Journal of Business Ethics*, 67(2), 107–123. https://doi.org/10.1007/s10551-006-9007-7

Baker, W., Benmark, G., Chopra, M., & Kohli, S. (2018). Master the Challenges of Multichannel Pricing. *MIT Sloan Management Review*, 60(1), 1-5.

Banaji, S., & Buckingham, D. (2009). THE CIVIC SELL. Information, *Communication & Society*, 12(8), 1197–1223. https://doi.org/10.1080/13691180802687621

Bolivar, A. (2005). El lugar de la ética profesional en la formación universitaria. *Revista Mexicana de Investigación Educativa*, 10, 93–123. Retrieved online from http://www.comie.org.mx/documentos/rmie/v10/n24/pdf/rmiev10n24scB06n01es.pdf

Boyd, D., & Crawford, K. (2012). Critical questions for big data - Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication and Society, 15(5), 662–679.

Buxarrais, M. R., & Esteban, F. (2004). el aprendizaje ético y la formación univeristaria: Más allá de la casualidad. Teoría de La Educación, 16(May), 91–108. https://gredos.usal.es/handle/10366/71927

Calo, R., & Rosenblat, A. (2017). The taking economy: Uber, information, and power. *Columbia Law Review*, 117, 1623. Retrieved online from https://columbialawreview.org/content/the-taking-economy-uber-information-and-power/

Castells, M. (1998). La era de la información: economía, sociedad y cultura Volumen III. Alianza.

Columbia Public Health (2021). Content Analysis. Retrieved online from https://www.publichealth.columbia.edu/research/population-health-methods/content-analysis

Couldry, N., & Turow, J. (2014). Advertising, big data, and the clearance of the public realm: Marketers' new approaches to the content subsidy. International Journal of Communication, 8(1), 1710–1726. http://repository.upenn.edu/asc_papers/413

Crawford, K., Miltner, K., & Gray, M. L. (2014). Critiquing Big Data: Politics, Ethics, Epistemology. Special Section Introduction. *International Journal of Communication* 2 (Vol. 8, Issue 1). https://doi.org/10.1080/10511259900084631

Crittenden, V., & Crittenden, W. (2015). Digital and social media marketing in business education: Implications for the marketing curriculum. *Journal of Marketing Education*, 37(2), 71–75. https://doi.org/10.1177/0273475315588111

Favaretto, M., De Clercq, E., & Elger, B. S. (2019). Big Data and discrimination: perils, promises and solutions. A systematic review. *Journal of Big Data*, 6(1). https://doi.org/10.1186/s40537-019-0177-4

Fieser, James: Ethic. Disponible en: Internet Encyclopidia of Philosophy (IEP). Recuperado de: https://iep.utm.edu/ethic/

Fourali, C. (2009). Developing world-class social marketing standards: A step in the right direction for a more socially responsible marketing profession. *Social Marketing Quarterly* (Vol. 15, Issue 2, pp. 14–24). https://doi.org/10.1080/15245000902957334

González, C. & Wangenaar, R. (2003). Tuning educational structures in Europe. España: Universidad de Deusto.

Google (2020). Dynamic Pricing Model For Online Advertising. Google Patent. Retrieved from https://patents.google.com/patent/US20110166927A1/en

Goutam, D. & Krishnendranath, M. (2017). A literature review on dynamic pricing of electricity, *Journal of the Operational Research Society*, 68:10, 1131-1145. https://doi.org/10.1057/s41274-016-0149-4

Gozálvez Pérez, V., García-Ruiz, R., & Aguaded-Gómez, J. I. (2014). La universidad como espacio de aprendizaje ético. In rieoei.org (Vol. 79, Issue 1).

Hannak, A., Soeller, G., Lazer, D., Mislove, A., & Wilson, C. (2014). Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 conference on internet measurement conference, November 2014* (pp. 305-318).

Harrigan, P., & Hulbert, B. (2011). How can marketing academics serve marketing practice? the new marketing DNA as a model for marketing education. *Journal of Marketing Education*, 33(3), 253–272. https://doi.org/10.1177/0273475311420234

Islam, G. (2021). Business Ethics and Quantification: Towards an Ethics of Numbers. *Journal of Business Ethics*, Taylor 2020. https://doi.org/10.1007/s10551-020-04694-z

Jacobson, J., Gruzd, A., & Hernández-García, Á. (2020). Social media marketing: Who is watching the watchers? *Journal of Retailing and Consumer Services*, 53, 1–12. https://doi.org/10.1016/j.jretconser.2019.03.001

Kerin R.A., Hartley S.W. (2019). *Marketing*. McGraw Hill Education. 14th Edition.

Kingsley-Hughes, A. (2012). Mac Users Have Money to Spare, Says Orbitz. Forbes, Jun. 26. Retrieved online from https://www.forbes.com/sites/adriankingsleyhughes/2012/06/26/mac-users-have-money-to-spare-says-orbitz/?sh=23b538952d59

Kosoff, M. (2016). You're more likely to accept Uber's surge pricing when your phone's about to die. Uber's head of economic research tells all. *Vanity Fair*, May 2016. Retrieved online from https://www.vanityfair.com/news/2016/05/uber-surge-pricing-low-phone-battery

Kotler P. & Amstrong G. (2018). *Principios de marketing*. Pearson (17ª edición). Madrid.

Lekakis, E. J. (2014). ICTs and ethical consumption: The political and market futures of fair trade. *Futures*, 62, 164–172. https://doi.org/10.1016/j.futures.2014.04.005

Liozu, S. M. (2019). Make pricing power a strategic priority for your business. *Business Horizons*, 62(1), 117-128.

Lobschat, L., Mueller, B., Eggers, F., Brandimarte, L., Diefenbach, S., Kroschke, M., & Wirtz, J. (2021). Corporate digital responsibility. *Journal of Business Research*, 122(July 2018), 875–888. https://doi.org/10.1016/j.jbusres.2019.10.006

Manuel Larrán, J., Andrades Peña, F. J., & Muriel de los Reyes, M. J. (2014). La responsabilidad social corporativa en las titulaciones de empresa y marketing ofertadas por las universidades españolas.

Esic Market Economics and Business Journal, 45(1), 121–146. https://www.esic.edu/documentos/revistas/esicmk/140217_163642_E.pdf

Marshall G.W. & Johnston M. W. (2015) *Marketing Management*. McGraw Hill Education. 2nd Editon.

Martin N. (2019). Uber Charges More If They Think You're Willing To Pay More. Forbes, Mar 30, 2019,12:58pm. Retrieved from https://www.forbes.com/sites/nicolemartin1/2019/03/30/uber-charges-more-if-they-think-youre-willing-to-pay-more/?sh=6a52775b7365

Mattioli D. (2012). On Orbitz, Mac Users Steered to Pricier Hotels. *The Wall Street Journal*, Ag. 23, 2012. Retrieved online from https://www.wsj.com/articles/SB10001424052702304458604577488822667325882

Maximilien, E. M., & Singh, M. P. (2005). Agent-based trust model involving multiple qualities. *Proceedings of the International Conference on Autonomous Agents*, 653–660. https://doi.org/10.1145/1082473.1082552

Montes, D. A., & Suárez, C. I. (2016). La formación docente universitaria: Claves formativas de universidades españolas. *Revista Electronica de Investigacion Educativa*, 18(3), 53–61. http://www.scielo.org.mx.sabidi.urv.cat/scielo.php?pid=S1607-40412016000300004&script=sci_arttext

Moscoso Pozo, M. M., Ruiz Zambrano, A., & Aragundi García, L. I. (2017). La educación y la formación profesional del marketing. Su abordaje desde una perspectiva ética compleja. *Didasc@lia: Didáctica y Educación*, VIII (4), 1–8.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 376, Issue 2133). https://doi.org/10.1098/rsta.2018.0089

OCDE (2019): Estrategia de competencias de la OCDE 2019. Competencias para construir un futuro mejor. Fundación Santillana.

OCDE. (2002). Definition and Selection of Competences (DESECO). Retrieved online from https://www.oecd.org/education/skills-beyond-school/definitionandselectionofcompetenciesdeseco.htm

Pemberton, J. D., Stonehouse, G. H., & Barber, C. E. (2001). Competing with CRS-generated information in the airline industry. *Journal of Strategic Information Systems*, 10(1), 59-76.

Schindler R.M. (2012). *Pricing Strategies: A Marketing Approach*. Sage publictions.

Schmid, B. (1994). Electronic markets in tourism. In *Information and Communications Technologies in Tourism* (pp. 1-8). Springer, Vienna.

Sharma, G., & Lijuan, W. (2014). Ethical perspectives on e-commerce: An empirical investigation. *Internet Research*, 24(4), 414–435. https://doi.org/10.1108/IntR-07-2013-0162

Simchi-Levi, D. (2017). The new frontier of price optimization. *MIT Sloan Management Review*, 59(1), 22.

Spanish Ministry of Science, Innovation and Universities (20 April 2021). Retrieved online from https://www.educacion.gob.es/notasdecorte/busquedaSimple.action

Tomlinson, E. C., & Mayer, R. C. (2009). The role of causal attribution dimensions in trust repair. *Academy of Management Review* (Vol. 34, Issue 1, pp. 85–104). Academy of Management. https://doi.org/10.5465/AMR.2009.35713291

Veeraraghavan S. (2016). The Price Is Pliant: The Risks and Rewards of Dynamic Pricing. Interview at Knowledge@Wharton, University of Pensilvania. Retrieved from https://knowledge.wharton.upenn.edu/article/price-pliant-considering-risks-rewards-dynamic-pricing

6. Open Track

EXPLORING THE JAPANESE GREY DIGITAL DIVIDE IN THE PANDEMIC ERA

Simon Rogerson, Tatsuya Yamazaki, Yohko Orito, Kiyoshi Murata

De Montfort University (UK), University of Toyama (Japan), Ehime University (Japan), Meiji University (Japan)

srog@dmu.ac.uk; tatsuya@eco.u-toyama.ac.jp; orito.yohko.mm@ehime-u.ac.jp; kmurata@meiji.ac.jp

ABSTRACT

This paper discusses an empirical study undertaken of a sample of Japanese people across the digital divide, focusing on their perception of both connectivity and being informed as the pandemic unfolds. The aim is to identify common themes regarding how digital technology is used to support information and interaction during the pandemic. These are used to propose changes which might halt the grey digital divide becoming the grey digital chasm and improve support through fit-for-purpose digital technology to the most vulnerable in times of emergency. To achieve the aim, questionnaire surveys were conducted and 136 valid responses including ones from grey digital natives and outcasts were analysed. The results of the analysis demonstrate that the grey digital divide did not seem to exist in terms of the acquisition of information about COVID-19, and that both grey digital natives and outcasts preferred to receive such information via low- and no-tech media. However, Japanese grey digital outcasts may receive a low priority regarding COVID-19 vaccination which began in April 2021, due to Japanese local governments' adopting online systems to administer the vaccination programme.

INTRODUCTION

The world has changed since the pandemic. Online has become the accepted first channel of choice for communication and social interaction. Or, perhaps the world at large had no choice but to accept that online was the only viable method of ongoing social interaction and communication. Whichever is the reason, the fact remains that the gap has widened between those digitally included and those digitally excluded. The digital divide is rapidly becoming a digital chasm. The digital divide has come about through increasing global digital technological dependency. It is a social divide which sits upon a complex foundation of poverty, education, gender, age, status, location, and mental and physical faculty.

In times of global emergency, such as the COVID-19 pandemic, social divides are extremely harmful. It is morally unacceptable to ignore the risk of harm and to fail to provide essential support to the vulnerable. There has been a tendency by some governments and some service providers to be reactive rather than proactive in their approach. Whilst reactive action in a state of emergency is appropriate so also is proactive action which has a preventative function. As society's voices of concern rise, mirroring the increasing suffering of the most vulnerable, so those in a position to alleviate the suffering react. Proactive action is needed both to minimise the chance of short term suffering and to instigate long term strategies to remove barriers and promote wellbeing.

The relationship between the COVID-19 pandemic and digital technology is worthy of study because digital technology has become the de facto standard for communication of alerts, advice and regulation announcements. Digital outcasts are left out of this loop and as such could be disenfranchised, disheartened and damaged. This paper discusses an empirical study undertaken of a

sample of people across the digital divide, focusing on their perception of both connectivity and being informed as the pandemic unfolds. It is generally accepted that those of 65 years and over are most at risk in the pandemic; the older a person is the greater the risk (see, for example, Kang & Jung, 2020 and Signorelli & Odone, 2020). For this group, termed *the elderly*, the digital divide has likely increased the risk of deteriorating health and tragically increasing the likelihood of death. The paper drills down into one specific population – Japan, where digital technology has pervaded society, the ageing of the population is being accelerated at a pace exceeding the rest of the world, and the number of confirmed COVID-19 cases remains at the lowest level among developed countries although strong measures to contain the spread of the novel coronavirus such as a national lockdown have not been taken. The elderly are placed in the context of a wider demographic through the chosen sample.

The aim is to identify common themes regarding how digital technology is used to support information and interaction during the pandemic. These are used to propose changes which might halt the grey digital divide becoming the grey digital chasm and improve support through fit-for-purpose digital technology to the most vulnerable in times of emergency.

WORLD VIEW

It is necessary to place the empirical study into a global context so that perspective and relevance can be ascertained. Table 1 shows the global and Japanese situations between March 2020 and March 2021.

Table 1. Pandemic comparisons over time.

date	confirmed cases				deaths				vaccine doses			
	World	% inc	Japan	% inc	World	% inc	Japan	% inc	World	% inc	Japan	% inc
07/03/2020	107,444		411		3,510		6					
07/08/2020	18,981,861	17,566.7	43,815	10,560.6	705,337	19,995.1	1,033	17,116.7				
07/11/2020	49,919,672	163.0	105,914	141.7	1,243,539	76.3	1,809	75.1				
07/03/2021	115,653,459	131.7	437,892	313.4	2,571,823	106.8	8,178	352.1	249,160,837		28,530	
31/03/2021	127,877,462	10.6	472,112	7.8	2,796,561	8.7	9,113	11.4	520,540,106	108.9	46,469	62.9
% against	1.628		0.374		0.036		0.007		6.627		0.037	
population	7,854,570,325		126,195,091		7,854,570,325		126,195,091		7,854,570,325		126,195,091	

Source: WHO COVID-19 dashboard at https://covid19.who.int/

The % Inc column shows the percentage increase from the previous date to the date in question. For example, between March 2020 and August 2020 there was a global 17,566.7% increase of confirmed cases and a Japanese 10,560.6% increase of confirmed cases. During March 2021, the percentage increase globally had fallen to 10.6% and in Japan, had fallen to 7.8%. Nearly 2.8 million people have died of COVID-19 of which nearly 10 thousand reside in Japan. Between 7 March 2021 and 31 March 2021 global vaccine doses increased by 108.9% to over 520 million people being vaccinated which represents 6.627% of the global population. In Japan for the same period, vaccine doses administered increased by 62.9% to over 46 thousand people which represents 0.037% of the country's population.

The large difference in death rates related to COVID-19 by age has been reported widely (see, from example, Signorelli & Odone, 2020, and Kang & Jung, 2020). The elderly are more likely to die of COVID-19 than the young. Omori et al (2020) have found that across Italy, Spain, and Japan, the age distributions of COVID-19 mortality show only small variation even though the number of deaths per

country shows large variation. This suggests that a study in one of these countries will have relevance to the other two and maybe beyond.

On 25 March 2020 it was reported by the BBC that one quarter of the world's population was living under some form of lockdown. Armitage & Nellums (2020) explain that self-isolation will disproportionally affect the elderly because of increased risk of cardiovascular, autoimmune, neurocognitive and mental health problems. It is estimated that 9% of the global population of 7.8 billion people are over 65 years of age[31]. Current access to the Internet stands at 62% of the global population[32] and of this 7% (0.3385 billion) are over 65 years of age[33]. This means that only 48% of the global population over the age of 65 years of age can be classified as grey digital natives. Therefore, the grey digital divide comprises 363.5 million digital outcasts. They are at particular risk during the COVID-19 pandemic because they become more isolated through their lack of communicative support by the authorities who tend to inform only through digital technology conduits. The global distribution of these outcasts will likely mirror disparities between developed and developing regions, urban and rural, rich and poor and literate and illiterate people. By way of illustration, out of the total global digital outcast population 27% reside in Africa, 31% in Southern Asia and 19% in Eastern Asia[34].

The global Internet economy has been defined as comprising three components: access provision – how we connect; service infrastructure – how we build and sustain the Internet; and Internet applications – how we communicate, share and innovate (Internet Society, 2020). During the pandemic, the performance of these components has been mixed. For example, video conferencing through portals such as Zoom, has provided excellent links for digital natives although "zoom burnout" has become a new phenomenon. However, broadband in rural areas has significantly reduced the ability for reaching out to the elderly in those areas.

DEFINITIONS

In this paper the following definitions are used (Rogerson, 2021).

- Demographic Profile: includes age, gender, ethnicity, faith, literacy and economic status.
- Digital Divide: the disparity in access, usage and benefit of any digital technology. On one side are those who are digitally included and on the other are those who are digitally excluded.
- Digital Native: the digitally literate, regardless of demographic profile, who use, and are somewhat dependent upon, digital technology.
- Digital Outcast: those, regardless of demographic profile, who are unable, for whatever reason, to access the benefits offered through the use of digital technologies.
- Elderly: anyone of 65 years and over. The elderly are often referred to as grey to differentiate them as population group; hence grey digital divide is now a commonly used term.

[31] statista at https://www.statista.com/statistics/265759/world-population-by-age-and-region/

[32] World Internet Usage and Population Statistics 2020 Year-Q2 Estimates at 20 June 2020 at https://www.Internetworldstats.com/stats.htm

[33] statista at https://www.statista.com/statistics/272365/age-distribution-of-Internet-users-worldwide/

[34] https://thenextweb.com/growth-quarters/2020/01/30/digital-trends-2020-every-single-stat-you-need-to-know-about-the-internet/

- Rogerson (ibid) explains that in any investigation of digital technology usage by the elderly during the pandemic, it is appropriate to use new terms Grey Digital Native and Grey Digital Outcast to describe the elderly positioned either side of the grey digital divide.

GREY DIGITAL DIVIDE TYPOLOGY

It is necessary to provide a more detailed description of the terms Grey Digital Native and Grey Digital Outcast. This will then enable the collected empirical data to be analysed using these descriptions. These detailed descriptions are derived by extending the user typology created by Birkland (2019) since this typology only covered the Native side of the Grey Digital Divide.

Birkland (2019) defines five types of Grey Digital Native:

- Enthusiast – is very positive about information and communication technology (ICT) and views it as a fun toy
- Practicalist – views ICT as a utility which enable specific tasks to be undertaken
- Socializer – uses ICT to extend interactions across intergenerational networks and communities
- Traditionalist – views ICT of the past as something to cherish rather than modern versions
- Guardian – views all ICT with caution and wants to act as a guardian against the downside of ICT usage.

To complete the topology five types of Grey Digital Outcast are suggested (Rogerson, 2020):

- Impoverished – having low economic status and/or limited ICT resources
- Isolated – geographically or socially remote from ICT users
- Illiterate – having a low level of general or technological knowledge
- Wary – being apprehensive of technology in general and ICT in particular
- Uninterested – seeing no purpose in accessing ICT rather than other channels and other activities

Birkland's user typology thus evolves into a Grey Digital Divide Typology as shown in Table 2 (Rogerson, 2020). It is this typological tool which can be used to analyse the empirical data. Circumstances and experiences could cause the elderly to move between types and across the divide. Indeed, Van Dijk (2005) explains that the elderly move between native and outcast depending upon mental, physical, financial and motivational circumstance. Given the typology it seems possible that an individual could reside in more than one type on one side of the divide.

Table 2. The grey digital divide typology.

	Grey Digital Natives (Birkland 2019)	**Grey Digital Outcasts (Rogerson 2020)**
1	Enthusiast	Impoverished
2	Practicalist	Isolated
3	Socializer	Illiterate
4	Traditionalist	Wary
5	Guardian	Uninterested

Level of digital technology usage is considered an important second dimension in the analysis. The use or non-use of digital technology is differentiated by a simple three-part classification: *no-tech* – print media, written letters and face to face dialogue; *low-tech* – television, radio and telephone; and *high-tech* – smartphone, social media and Internet (Rogerson, 2021).

JAPAN OVERVIEW

According to *worldometer* the population of Japan is 126.48 million with 92% living in urban areas and 8% living in rural areas. The life expectancy is 85.03 years. Japan has the world's largest percentage of elderly adults totalling 35.58 million (source: www.prb.org/countries-with-the-oldest-populations). As seen in Table 1, as at 31 March 2021, there have been 472,112 confirmed cases of COVID-19 resulting in 9,113 deaths.

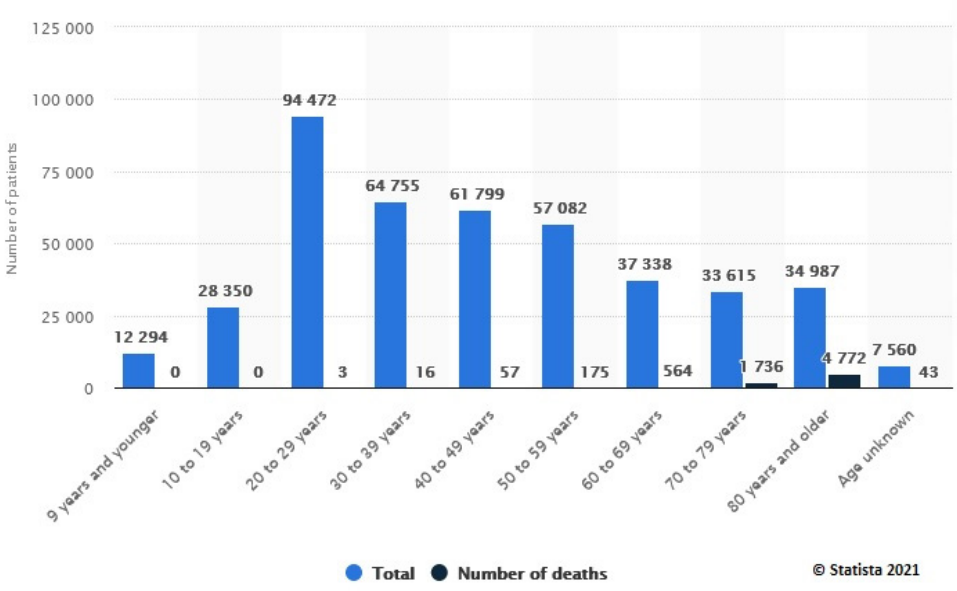
Figure 1 shows cases and deaths across the age ranges. This distribution of coronavirus disease (COVID-19) cases in Japan as of March 3, 2021, shows that the highest number of patients were aged 20 to 29 years old, with a total of around 94.5 thousand cases. The highest number of deaths is among the patients aged 80 years and older at about 4.7 thousand cases. The Japanese health ministry announced on March 5 that there was a total of around 436.7 thousand confirmed cases of COVID-19 in Japan. This data shows that those aged over 60 years account for 24% (105,940) of reported cases yet they account for 96% (6,508) of reported COVID-19 related deaths. Those over 80 years of age accounted for 64.8% of deaths. This clearly demonstrates the primary need to protect the elderly through preventative medical care, appropriate communication and social support.

Statista (2020) reports that Japan is fourth highest internet user in East Asia with 117 million users but the distribution of this population was affected by age, region, and income level. The Internet in Japan usually comes unbundled, meaning that customers in general have contracts both with a line provider and an internet service provider (ISP). Okabe (2020a) suggests that there are many elderly Japanese who are unfamiliar with using digital technologies which lack the traditional and familiar means of identification such as the personal stamps (*inkan* 印鑑) and common seals (*kōin* 公印). Furthermore, Okabe (2020b) suggests many elderly and impaired people of becoming excluded from retail which is increasingly moving online as these people cannot use the digital equipment and are unfamiliar with online shopping techniques. It seems there needs to be improved digital social scaffolding for the elderly. The White Paper Information and Communication in Japan (MIC, 2020) reported that Internet usage rate is 89.8% in 2019 for the whole population (79.8% in 2018), 90.5% for 60-69 year olds (76.6% in 2018), 74.2% for 70-79 year olds (51.0% in 2018), and 57.5% for those 80 years old and over (21.5% in 2018). This suggests that there remains a grey digital divide albeit diminished and this becomes more significant in the over 80 year olds where around 65% of COVID-related deaths occur. This implies that there exists a reasonable population of grey digital natives and grey digital outcasts from which to sample.

SURVEY METHOD

To investigate elderly Japanese people's attitudes and behaviour towards COVID-19 in relation to digital technologies in particular, questionnaire surveys were conducted online – using Google Forms – as well as offline – using a pencil-and-paper questionnaire. The survey was conducted in Tokyo and four local cities in Japan (Bizen, Chiryu, Matsuyama and Takaoka) from August to December 2020,

Figure 1. Patient profile of COVID-19 cases in Japan as of March 2021, by age group.



Source: https://www.statista.com/statistics/1105162/japan-patients-detail-novel-coronavirus-covid-19-cases-by-age-and-gender/

with the support of those experts who were engaged in welfare service for the elderly in Japan or local community support. The original English questionnaire, which was qualitative-based using seven open-ended questions, was developed by one of the authors (Rogerson), and was translated into Japanese by the rest. Based on the suggestion given by one of the experts, response alternatives were set in three of the seven open-ended questions to ease strain on elderly respondents when responding to the questionnaire. Out of 141 responses, 136 were deemed to be valid of which 32 were provided online. The attributes of respondents are shown in Table 3. There are 67 digital outcasts of which 59 are grey digital outcasts and 69 digital natives of which 34 are grey digital natives. That more than 60% of respondents above 65 were grey digital outcasts demonstrates the collected data was suitable for our research purpose. The data was analysed statistically using chi-squared tests and the User Local text mining tool[35].

Table 3. Attributes of respondents.

	Internet usage		
Age group	**Internet users**	**Non-Internet users**	**Total**
Above 65	34	59	93
Between 60 and 65	22	3	25
Below 60	13	5	18
Total	69	67	136

[35] https://textmining.userlocal.jp/

SURVEY FINDINGS

Perceived source of information about COVID-19 for the public

The survey results indicate that most of respondents considered the public's main sources of information about COVID-19 were TV news and newspapers as shown in Figure 2. It is statistically confirmed that this tendency did not vary between users and non-users of the Internet, as well as across age groups. On the other hand, as Table 4 shows, more than a half of respondents above 65 felt tabloid TV shows (55 out of 93) and family members or friends (49 out of 93) were information sources about the disease, whereas significantly a lower percentage (less than 30%) of respondents at age 65 years or younger considered they (12 and 11 out of 43, respectively) were important information sources for the public. While online news sites are considered a good information source for the public by more than a half of respondents at age 65 years or younger, less than 20% of respondents above 65 felt this, showing a significant difference at 1% level between age groups. Official government announcements made online were not recognised a helpful information source for the public by most respondents regardless of age group or Internet usage (p=.3761).

Figure 2. Perceived source of information about COVID-19 for the public (N=136; multiple answers allowed).

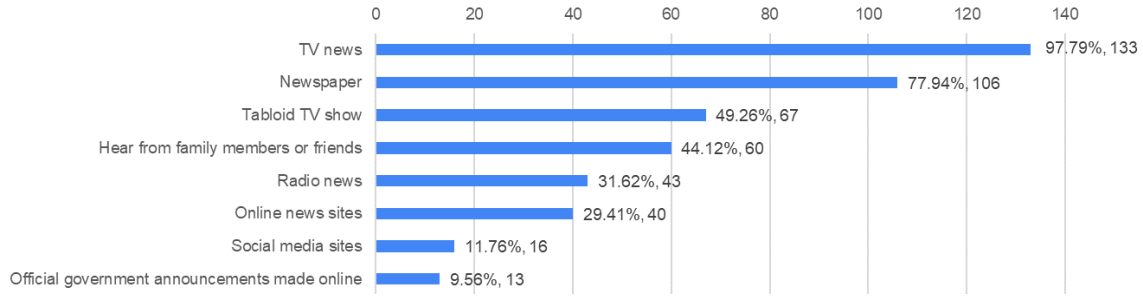


Table 4. Perceived source of information about COVID-19 for the public by age groups (multiple answers allowed).

	Above 65 (93)		At age 65 years or younger (43)			
How is information about COVID-19 made available to the public?	**Number**	**%**	**Number**	**%**	***p*-value**	
TV news	91	97.85%	42	97.67%	.9923	
Newspaper	79	84.95%	27	62.79%	.1736	
Tabloid TV show	55	59.14%	12	27.91%	.0158	*
Hear from family members or friends	49	52.69%	11	25.58%	.0269	*
Radio news	39	41.94%	4	9.30%	.0016	**
Online news sites	18	19.35%	22	51.16%	.0015	**
Social media sites	6	6.45%	10	23.26%	.0079	**
Official government announcements made online	11	11.83%	2	4.65%	.2081	

As shown in Table 5, TV news, newspapers, tabloid TV shows and family members or friends were perceived to be the public's source of information about COVID-19 by at least half of respondents above 65, regardless of whether they are Internet users or not. There was a significant difference in respondents' recognition concerning online news sites as an information source about the disease for the public between Internet users and non-users at 1% level. A large majority of respondents above 65, regardless of their Internet usage, did not feel official government announcements made online to be a COVID-19 information source.

Table 5. Source of information about COVID-19 for the public perceived by respondents above 65 (multiple answers allowed)

	Internet users (34)		Non-Internet users (59)			
How is information about COVID-19 made available to the public?	**Number**	**%**	**Number**	**%**	**Total**	***p*-value**
TV news	33	97.06%	58	98.31%	91	.9533
Newspaper	28	82.35%	51	86.44%	79	.8368
Tabloid TV show	17	50.00%	38	64.41%	55	.3843
Hear from family members or friends	19	55.88%	30	50.85%	49	.7473
Radio news	15	44.12%	24	40.68%	39	.8051
Online news sites	14	41.18%	4	6.78%	18	.0003 **
Social media sites	5	14.71%	1	1.69%	6	.0174 *
Official government announcements made online	3	8.82%	8	13.56%	11	.5225

Respondents' preferred source of information about COVID-19

As with respondents' perception of information source for the public, TV news and newspapers were the sources of information about COVID-19 most preferred by respondents (Figure 3). However, there were statistically significant differences in the preference tendencies between age groups. Respondents above 65 tended to prefer to access information about COVID-19 via TV news, newspapers, Tabloid TV shows, family members or friends and radio news more than respondents at age 65 years or younger at 1% level, as shown in Table 6. On the other hand, online news sites were preferred information sources for respondents at age 65 years or younger more than for respondents above 65 at 1% level. However, a large majority of respondents did not prefer to gain information about the disease from official government announcements made online regardless of their age.

Figure 3. Respondents' preferred source of information about COVID-19 (N=136; multiple answers allowed)

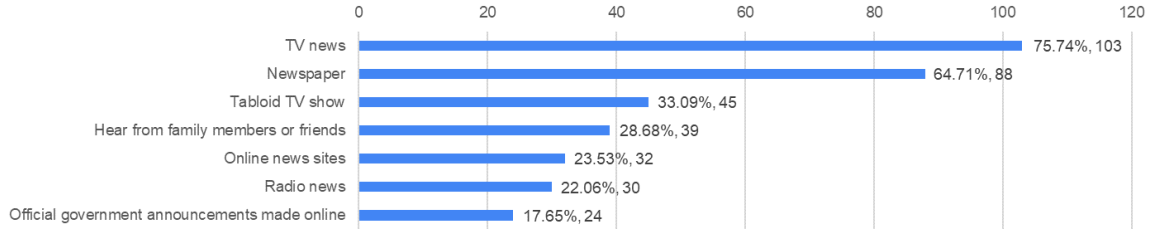


Table 6. Respondents' preferred source of information about COVID-19 by age groups (multiple answers allowed)

	Above 65 (93)		At age 65 years or younger (43)		
What is your preferred way of receiving information about COVID-19?	**Number**	**%**	**Number**	**%**	***p*-value**
TV news	83	89.25%	20	46.51%	.0077 **
Newspaper	75	80.65%	13	30.23%	.0007 **
Tabloid TV show	40	43.01%	5	11.63%	.0031 **
Hear from family members or friends	38	40.86%	1	2.33%	.0001 **
Online news sites	15	16.13%	17	39.53%	.0089 **
Radio news	29	31.18%	1	2.33%	.0009 **
Official government announcements made online	13	13.98%	11	25.58%	.1342

As shown in Table 7, a statistically significant difference in above-65-respondents' preference to online news sites as sources of information about COVID-19 was found at 1% level. However, Internet usage by respondents above 65 tended not to affect significantly their preferred ways of receiving information about the disease.

The outcomes of text mining of responses to the open-ended question about reasons for respondents' preferred ways of receiving information about COVID-9 suggest that they eagerly desired to get correct, unbiased and trustworthy information, whenever they want and in a convenient way (Figure 4). 53 respondents, of which 6 were non-Internet users, responded to this question, and many of them emphasised the importance of getting correct and unbiased information from trustworthy sources.

Typical comments from respondents were, "I want to get as correct information as possible", "I want to get information from various sources and judge autonomously" and "I'd like to get information from trustworthy people", despite the differences in their preferred ways of getting information about the disease. The ease and timeliness of information acquisition were also stressed by many. A respondent using the Internet commented, "Online news sites provide us with news most promptly. We can access to them anytime and anywhere. This is convenient", whereas another who was a non-Internet user mentioned, "[TV news is] the easiest way to get information". These attitudes may be reflected in the survey results that respondents' main sources of information about the disease were TV news and newspapers and even grey digital natives tended not to search actively such information online accessing, for example, dedicated websites for COVID-9 such as the Novel Coronavirus Website of the Ministry of Health, Labour and Welfare[36]. However, given the fully commercialised TV media in Japan and its low rank of freedom of the press (67th according to the 2021 world freedom press index issued by the Reporters without Borders[37]), this may show that Japanese grey digital outcasts, as well as natives, receive distorted information about the pandemic: their sense of fear of COVID-19 may be unnecessarily promoted, and what they can know about the disease may be controlled by the government.

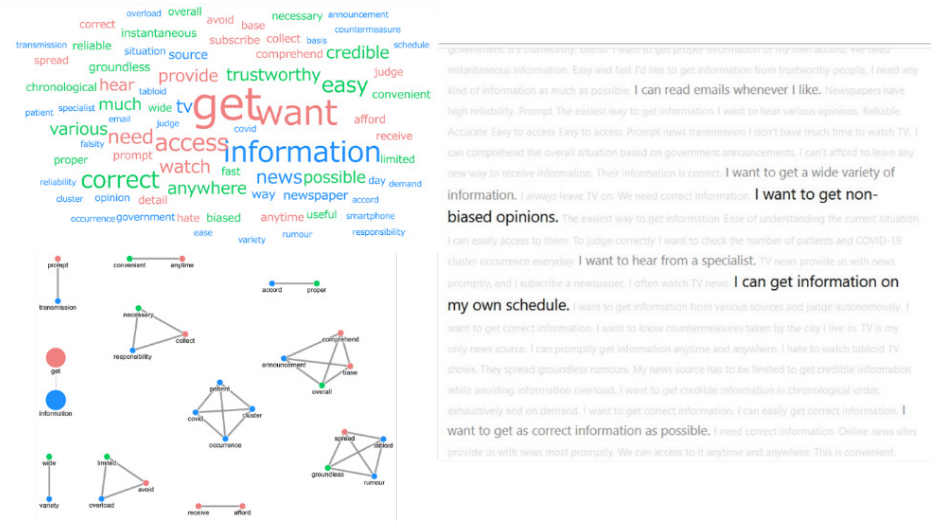
[36] https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000164708_00001.html

[37] https://rsf.org/en/japan

Table 7. Preferred source of information about COVID-19 for respondents above 65 (multiple answers allowed).

	Internet users (34)		Non-Internet users (59)			
What is your preferred way of receiving information about COVID-19?	**Number**	**%**	**Number**	**%**	**Total**	***p*-value**
TV news	26	76.47%	57	96.61%	83	.3221
Newspaper	25	73.53%	50	84.75%	75	.5619
Tabloid TV show	11	32.35%	29	49.15%	40	.2342
Hear from family members or friends	9	26.47%	29	49.15%	38	.0994
Online news sites	13	38.24%	2	3.39%	15	.0001 **
Radio news	10	29.41%	19	32.20%	29	.8164
Official government announcements made online	6	17.65%	7	11.86%	13	.4726

Figure 4. Text mining outcomes (Reason for preferred ways of receiving information about COVID-19).



SYNTHESIS

To reveal the characteristics of attitudes and behaviour of Japanese grey digital natives and outcasts based on the survey results, we focus only on respondents above 65 and use the grey digital divide typology and the level of digital technology usage. Hereafter, grey digital natives mean respondents above 65 who used the Internet, and grey digital outcasts mean those who did not use the Internet. Based on their responses to the question on the purpose of Internet usage, 33 grey digital natives (one of the respondents did not answer this question) can be classified as follows: 13 enthusiasts, 6 practicalists, 13 socializers and 1 traditionalist. However, as Table 8 shows, a large majority of high-tech users – even enthusiasts – preferred to get information about COVID-19 from low-tech – TV news – and no-tech media – newspapers.

It is hard to imagine that grey digital outcasts fell into the classification of impoverished or isolated, given their residential areas and the fact that paper-based survey data were collected by the experts mentioned in Section 6 – all of them were ICT users. It is possible that grey digital outcasts were illiterate, wary or uninterested. As shown in Table 9, just 7 out of 59 grey digital outcasts preferred to acquire information about COVID-19 online demonstrating a large majority of them were would-be low-tech users, whereas there was no one who preferred to receive information only from no-tech media.

Table 8. Preferred source of information about COVID-19 for grey digital natives (multiple answers allowed).

	Types of grey digital natives							
	Enthusiast (13)		Practicalist (6)		Socializer (13)		Traditionalist (1)	
What is your preferred way of receiving information about COVID-19?	Number	%	Number	%	Number	%	Number	%
TV news	11	84.62%	4	66.67%	10	76.92%	0	0.00%
Newspaper	11	84.62%	3	50.00%	10	76.92%	0	0.00%
Online news sites	6	46.15%	3	50.00%	4	30.77%	0	0.00%
Tabloid TV show	4	30.77%	2	33.33%	5	38.46%	0	0.00%
Radio news	3	23.08%	2	33.33%	5	38.46%	0	0.00%
Hear from family members or friends	4	30.77%	1	16.67%	4	30.77%	0	0.00%
Official government announcements made online	3	23.08%	2	33.33%	1	7.69%	0	0.00%
Social media sites	1	7.69%	0	0.00%	1	7.69%	1	100.00%
Magazine	0	0.00%	1	16.67%	2	15.38%	0	0.00%
City office's website	1	7.69%	0	0.00%	0	0.00%	0	0.00%
Blogs and websites personally run	0	0.00%	0	0.00%	1	7.69%	0	0.00%
Local city news	1	7.69%	0	0.00%	0	0.00%	0	0.00%
Relevant books	1	7.69%	0	0.00%	0	0.00%	0	0.00%

Table 9. Preferred source of information about COVID-19 for grey digital outcasts (multiple answers allowed)

	Would-be low-tech user (52)		Would-be high-tech user (7)	
What is your preferred way of receiving information about COVID-19?	Number	%	Number	%
TV news	51	98.08%	6	85.71%
Newspaper	43	82.69%	7	100.00%
Hear from family members or friends	23	44.23%	6	85.71%
Tabloid TV show	23	44.23%	6	85.71%
Radio news	15	28.85%	4	57.14%
Official government announcements made online	0	0.00%	7	100.00%
Magazine	3	5.77%	3	42.86%
Online news sites	0	0.00%	2	28.57%
Information from municipal office	1	1.92%	0	0.00%
Blogs and websites personally run	0	0.00%	1	14.29%

CONCLUSION

The results of our questionnaire surveys suggest that the phenomenon of grey digital divide does exist in Japan in the sense that there was a significant number of Interne users among respondents above 65 (34 out of 93) whereas more than 60% of them (59) were non-Internet users. However, the grey digital divide did not seem to exist in terms of the acquisition of information about COVID-19 at least as of December 2020 when the surveys were completed.

More than 60% of grey digital natives (21 out of 33) used the Internet for communicating with others by email (18), instant messenger (15) and/or toll-free phone call (9), and nearly a half of them (16) for online search. Though they could potentially use digital technology to access various information about the disease through searching and interacting online, Japanese grey digital natives showed their reluctance to accessing such information, preferring to depend on low- and no-tech media.

On the other hand, we can make a good guess that grey digital outcasts communicate with others using phone calls, written letters and/or face to face dialogue. In fact, nearly a half of them (29 out of 59) preferred to receive information about COVID-19 through hearing from family members or friends, whereas less than 30% of digital natives (9 out of 34) preferred this route. Nearly 90% of grey digital outcasts (52 out of 59) were would-be low-tech users, meaning that they did not have an intention to use the Internet to get information about the disease even during the pandemic era. In Japan, however, vaccination against COVID-19 has given to a greying generation since April 2021 at a really sluggish pace, and many of local governments that are in charge of the vaccination adopt online systems which allow people to make reservations for the vaccination via smartphone or PC. This may result in making the latent grey digital divide a reality, leading to grey digital outcasts disadvantaged and even penalised within the vaccination programme.

ACKNOWLEDGEMENTS

The authors appreciate the cooperation for conducting the surveys provided by Professor Shizuka Suzuki, Ms. Qi Zhang and Ms. Haruka Suzuki of Ehime University, Mr. Masashi Kanegae of the Japan Support Center for Activity and Research for Older Persons, Mr. Hiroyasu Nakanishi of the Bizen Katakami Mutual Support Committee and Mr. Kazuhiko Takeichi of the Japan Senior Citizens' Council.

KEYWORDS: digital divide, digital outcast, COVID-19, Japanese elderly.

REFERENCES

Armitage, R. & Nellums, L. B. (2020). COVID-19 and the consequences of isolating the elderly. *Lancet Public Health*, *5*(5), p. e256.

Birkland, J. (2019). *Gerontechnology: Understanding Older Adult Information and Communication Technology Use*. Emerald Publishing.

Internet Society (2019). *Global Internet Report: The Forces Shaping Our Digital Future*. Available at: https://future.Internetsociety.org/2019/wp-content/uploads/sites/2/2019/04/InternetSociety-GlobalInternetReport-ConsolidationintheInternetEconomy.pdf

Kang, S.J. & Jung, S.I. (2020). Age-related morbidity and mortality among patients with COVID-19. *Infection & chemotherapy*, *52*(2), p. 154.

MIC (2020) *White Paper on Information and Communications in Japan*. Ministry of Internal Affairs and Communications, Japan, August.

Natale, F., Ghio, D., Tarchi, D., Goujon, A. & Conte, A. (2020). *COVID-19 Cases and Case Fatality Rate by age*. European Commisssion, 4 May. Available at: https://knowledge4policy.ec.europa.eu/publication/covid-19-cases-case-fatality-rate-age_en

Okabe, M. (2020a) *Preserve the hanko tradition.* Japan Times, 27 March. Available at: https://www.japantimes.co.jp/opinion/2020/03/27/reader-mail/preserve-hanko-tradition/

Okabe, M. (2020b) *Online shopping needs some improvement to be more inclusive.* Japan Times, 10 July. Available at: https://www.japantimes.co.jp/opinion/2020/07/10/reader-mail/online-shopping-needs-improvements-inclusive/

Omori, R., Matsuyama, R. & Nakata, Y. (2020). The age distribution of mortality from novel coronavirus disease (COVID-19) suggests no large difference of susceptibility by age. *Scientific reports, 10*(1), pp. 1-9.

Rogerson, S. (2020). *The Digital Divide: Grey Digital Outcasts and COVID-19*. Science Group, Buxton and District U3A, 20 November.

Rogerson, S. (2021). Grey Digital Outcasts and COVID-19. *American Behavioral Scientist*. forthcoming.

Signorelli, C. and Odone, A. (2020). Age-specific COVID-19 case-fatality rate: no evidence of changes over time. *International Journal of Public Health, 65*(8), pp. 1435-1436.

Statista (2020). *Internet usage in Japan - statistics & facts*. Statista Research Department, Aug 6. Available at: https://www.statista.com/topics/2361/internet-usage-in-japan/

Van Dijk, J.A. (2005). *The deepening divide: Inequality in the information society*. Sage Publications.

RESPONSIBLE PUBLIC ENGAGEMENT AT TERRITORIAL LEVEL: CORE DIMENSIONS AND MEANS FOR IMPLEMENTATION

Maria Michali, Thalia Kallia, Daniele Mezzana, George Eleftherakis

South-East European Research Centre (Greece), South-East European Research Centre (Greece), Knowledge & Innovation (Italy), CITY College, University of York Europe Campus (Greece)

mmichali@seerc.org; efkallia@seerc.org; mezzana@knowledge-innovation.org; eleftherakis@york.citycollege.eu

ABSTRACT

Responsible Research and Innovation (RRI) appeared as a policy concept acknowledged by the European Commission (EC). Over the last decade, RRI and its six keys form the basis for transformational initiatives at national, institutional and territorial levels. Particularly the Public Engagement (PE) key seeks to democratize Science, Technology and Innovation (STI); a key aspect in this democratic restructuring is the meaningful engagement of civil society and interested stakeholders in STI and research processes, and the consideration of their values and concerns. The present paper places the focus on RRI and PE application at territorial level (Territorial RRI framework). Territorial PE initiatives address a common STI-related territorial stake and engage the public through various methods –thus abiding by the public's local ethical codes and standards. By critically examining five PE-related territorial applications in five different territories, this paper aims to draw the attention to the conceptual underpinnings of territorial PE and to its practical application –as indicated by the experiences examined. The concluding arguments indicate some core conceptual traits of territorial RRI and PE. Then, they spell out specific elements contributing to the effective application of territorial PE for integrating public ethics concerns in STI processes and transformational agendas. Similarly, a few limitations, which should not be overlooked during territorial PE implementation, are outlined. The paper's evidence-based and concluding observations can finally provide valuable input for creating a new social (and ethical) contract between science and society, and for mitigating the exclusive dominance of the technocratic elite at territorial agendas.

INTRODUCTION

Responsible Research and Innovation (RRI) is a multi-dimensional concept gaining advancing prominence within the final decade in the ERA (European Research Area) and beyond. The EC (European Commission) characterizes RRI as "a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products" (Von Schomberg, 2011 as cited in Owen, Macnaghten, & Stilgoe, 2012, p. 753). Drawing on this EC-acknowledged definition of RRI, one explicitly comprehends that RRI principles and tenets (RRI keys) can highly contribute to aligning the agendas of STI (Science, Technology, Innovation) to society's needs and ethical concerns or values. A strong emphasis can be particularly placed on how the RRI key of Public Engagement (PE) contributes to achieving the aforementioned objectives. PE advocates, among others, towards: a) involving stakeholders so that innovations address societal needs, societal complexities and ethical problems (Taebi et al., 2014), b) engaging the public "before an issue or technology becomes controversial, when opinions become polarised and hardened and polices are

predetermined" (Cobb & Gano, 2012, p. 97) –thus making ethical questions acquire genuine meaning within the context of research, technology and development.

The six RRI keys, RRI principles, dimensions and related initiatives can be applied at a national, institutional and/or territorial level. With reference to the Territorial RRI framework, it can be described as shaping research and innovation (R&I) to support territory-making processes and new governance-making agendas (Caiati & Mezzana, 2019). It concurrently advocates that R&I and STI processes need to be responsive and 'response-able' to regional and societal needs, as well as to grand societal challenges (Fitjar, Benneworth, & Asheim, 2019). Consequently, territorial PE acquires new similar dimensions. It focuses on territory-based co-creation, and affects territory-making R&I agendas by combining scientific knowledge with the intimate knowledge of the territory's local actors (Caiati & Mezzana, 2019). Scientific knowledge is socially and ethically enhanced, since regions –owing to their scale– exhibit a better proximity towards social values and ethical concerns. Overall, experts can collaborate with citizens and practitioners, with the latter ones acquiring the license to express their ethical concerns towards new developments in the region, and subsequently infuse their own (otherwise tacit) ideas and moral standards to core regional transformational agendas.

The present paper discusses the application of Public Engagement (within the RRI umbrella) at the territorial level. It places an emphasis on the critical interpretation (dimensions, principles) of territorial PE, shifting the focus to engaging communities at various territorial scales (e.g. at a region, city, municipality). PE approaches, which can be applied at a territorial scale and within an upstream, midstream or downstream vantage point, are afterwards described in an evidence-based way. Territorial PE-related applications that are responsive to ethical and societal concerns relate to the following initiatives and specific PE methods (indicated in the parenthesis): BlueAp - Bologna Local Urban Environment Adaptation Plan for a Resilient City (Participatory design); Citizen science Lab of Leiden University (Science Shops); Sustainable Urban Mobility Plan Bremen 2025 (Scenario planning); Brainport Smart District (Living Labs), Energy vision Murau (Guiding Visions technique for Agenda Setting). The rationale of the above processes has after all been to critically describe the content of the aforementioned PE activities applied territorially, and subsequently stress some concluding elements that enhance the conceptual underpinnings and applications of territorial PE.

This holistic approach has the potential to lead to the successful 'operationalisation' of territorial PE for meeting society's ethical concerns towards R&I and technology. Valuable arguments are provided in this way towards enhancing: a) the concept of shared responsibility in innovation, and b) the opportunities for constructing new territorial and PE-driven R&I and technological agendas. These agendas ultimately promote a socially and ethically robust science by combining expert/scientific knowledge with local and practical experiential knowledge –the so-called building of a truly knowledge-based society (Steinhaus, 2013).

The present paper is structured as follows; firstly, the state-of-the art of RRI and PE is set, followed by their framing at the territorial context (Territorial RRI and PE). The focus then shifts exclusively on territorial PE. Its conceptual underpinnings, which among others entail addressing public ethical concerns towards STI, are primarily described. Immediately after, the five territorial PE applications are examined and valuable insights gained are spelled out. The paper's concluding remarks allude to experience-based arguments towards realistically and efficiently capitalizing on territorial PE for developing STI processes and agendas that are ethically responsive and democratic.

RRI, PE AND THEIR TERRITORIAL APPLICATION

The inability to harmonise scientific and technological knowledge with social and ethical responsibility is a challenge put forward several years ago (Mitcham, 2003; Stilgoe, Owen, & Macnaghten, 2013); as Innerarity (2013) argued, knowledge does not seize to be perceived as a product of experts, so as to then be 'open' to social guidance and turn into a social construct. At the same time, apart from the inability for 'openness' towards society, science seems to urge new manifestations of public hesitation and ethical concerns due to its increasing 'emancipation' (Mejlgaard et al., 2018). RRI came to manifest itself as the European Commission's response towards the aforementioned challenges. RRI emerged as a policy concept aiming to initiate transparent dialogues with society (de Saille, 2015) and achieve socio-technical collaboration in STI processes. This exact socio-technical integration in STI was further promoted by the EC by acknowledging the 'pillars' approach of RRI (Pellé & Reber 2015), and by promoting its application within various national, institutional and/or territorial initiatives. Based on the EC policy framework and the pillars approach, responsible and ethically accountable STI activities and outcomes should consider six key policy agendas, the so-called RRI keys: *Public Engagement*, *Open Access*, *Science Education*, *Gender*, *Ethics* and *Governance*. For ensuring the optimum outcomes, these RRI 'ingredients' can also be circumscribed by four core conditions: *anticipation*, *reflexivity*, *inclusion* and *responsiveness* (the so-called procedural approach). It can therefore be noticed that RRI has been attributed the ability to challenge the traditional social contract between science and society, and enhance new reconfigurations of actors and shared responsibilities in STI processes (Rip & Shelley-Egan, 2010).

As for the RRI key of Public Engagement (PE), the EC (2020) defines PE processes as "co-creating the future with the public and civil society organisations, and also bringing on board the widest possible diversity of people that would not normally interact on matters of science and technology". In other words, PE implies the establishment of participatory multi-actor interactions and exchanges, which can provide input to STI processes and policy agendas. Through such interactions, particularly the public can express ethics concerns towards emerging advancements and also create the space for the different values at stake to be expressed. As for the operationalisation of such expert-public interactions, these can broadly take place within three vantage points: *upstream*, *midstream* and *downstream* (Marschalek, 2017). The upstream perspective signifies that the public should be engaged at the early stage of research and technological development; it should contribute to answering *what* research questions or challenges the project/initiative should address. Midstream PE then signifies the stage of actual research and development, where the public can participate in the research process or provide input as to *how* the research could evolve. Finally, during downstream PE the public is usually asked about *whether* the outcomes and products of the STI processes should be adopted (and how). It should overall be highlighted that since such exchanges take place between both experts and public and the decisions are shaped collaboratively, PE exhibits an additional, beneficial effect. It manages to address a long-standing dilemma in R&I (Bucchi & Neresini, 2008, p.466) referring to the "technocratic option" (expert-driven decisions) or the "ethical option" (decisions driven by the individual users).

When applied at territorial level, RRI and PE acquire some new features. This application however is not automatic, in the sense that one must understand how RRI can be integrated into the territorial dynamics and with particular attention to the local actors and their concerns. This is an important question, since the original definition of RRI does not in itself have a spatial dimension (Fitjar, Benneworth, & Asheim, 2019). The territory is a human/social construction, which takes place by giving names and meanings to certain extensions of space, carrying out transformative material activities, elaborating and applying rules, transmitting all this over time, building a community. Concerning this, there are dynamics of de-territorialization (Paasi, 1998; Elden, 2005), that is, the loss

of social ties and control of the territory by the actors who live there (due to globalization, de-localization, unemployment, impacts of the environmental crisis, etc.) (Sassen, 2013). But reverse processes can also occur, of re-territorialization, to recover social bonds and identities. This can be done through territory-making practices (Dorstewitz, 2016) in different areas (energy, mobility, urban/rural development, services), implemented through forms of cooperation among local actors.

Based on a mapping of experiences that one can define as re-territorialization, it is possible to hypothesize that territory-making has at least three characteristics (Caiati & Mezzana, 2019): the development of a "territorial awareness", the activation of a "territorial mobilisation", the production of a "territorial change for governance". It is also possible to identify two components of territory-making *policies*: (a) the *"territorial orientation"*, which refers to what is intended to be done for and to be changed in the territory (e.g. re-rooting social and economic activities, empowering local actors, etc.); (b) the *"governance frameworks"*, which refer to the structured and recurring operating methods through which the territory-making process takes place (e.g. fostering a participative agenda setting system, launching knowledge co-creation platforms, etc.).

In this context, territorial RRI can be understood as the ability of R&I to respond to de-territorialization and contribute to re-territorialization, in terms of response-ability (Caiati & Mezzana, 2019). Various territorial RRI processes, while having a focus on re-territorialization, can shape the direction of R&I towards ethically desirable targets (at the European scale). Having said this, Territorial RRI can play a pivotal role, considering: (a) the RRI keys (how to use the RRI keys to open research and innovation to public concerns, to territory-making process and enhance the "territorial orientation"); (b) the RRI dimensions (how the four dimensions of anticipation, inclusiveness, responsiveness and reflexivity can be taken into account while using R&I for strengthening the territorial "governance frameworks", always in accordance to ethical accountability). With particular reference to the Public Engagement key and its territorial application, the focus is on a common territorial and R&I-related (or STI-related) stake, turned into reality through the cooperation among R&I actors and other key players, including individual citizens (Caiati & Mezzana, 2019).

TERRITORIAL PE: CONCEPTUAL BACKGROUND AND PRACTICAL APPLICATIONS

After setting the scene around RRI, PE and the new features they acquire when applied territorially, this section focuses explicitly on territorial PE. Its benefits in relation to including the public perspective in STI are firstly described based on previous literature. Section 3.2 afterwards describes five specific PE applications, which have taken place in five different territories and have employed five different PE methods.

Responsible PE: expression of ethics concerns towards sti

Shifting the focus on Territorial PE, it may be interpreted in terms of *territory-based co-creation*. Examples of territorial PE application can be, among others, the living labs, science-shops, and all forms of participatory design and citizen science. In more details, for achieving the objectives of a given territorial R&I policy, scientific knowledge should be combined with the knowledge of the territory. This knowledge, after all, stems out of its people and organisations that act in the territorial milieu, are bearers of the knowledge itself, and take into account the views and needs of the local community. In addition, territorial PE, due to its intrinsically participatory character, contributes to the identification, expression and sharing of ethical issues and concerns related to STI, and can foster a dialogue between different actors to better address these issues and concerns, and possibly identify ethical principles and codes tailored to the local context. As evident, the effects of territorial PE are enhanced by the

fact that the regional context around STI processes can be beneficial for effective knowledge acquisition and spillover (Laursen, Masciarelli, & Prencipe, 2012). The potential contribution of territorial PE is finally aligned to the EU and EC concerns. Various research programmes explicitly seek solutions to contemporary societal challenges (European Union, European Commission, & Directorate-General for Research and Innovation, 2013) and highlight necessity to address societal needs and ethical questions in research, development and technology.

PE practical examples at territorial level

As a follow-up to the description of the main conceptual features of Territorial PE, this section focuses on its practical application. Five PE-related applications have been selected to be critically described, so as to discuss a considerable range of PE methods contributing to the integration of public ethics concerns into STI processes, and promoting a responsible governance of science and technology. Based on a mapping of territorial RRI experiences (Caiati & Mezzana, 2019), the five PE methods listed in Table 1 are among the most frequent ones. Under this rationale, we have tried to detect through desk research concrete territorial initiatives employing these methods; by capitalising on the criterion of transparency and adequate public documentation, the target five initiatives that deal with a considerable challenge in the target territory have been sorted out. Prior to critically describing the application of PE in each of the five different initiatives, Table 1 outlines all five PE applications, in terms of territorial initiative, PE method and target territory.

Table 1. The five territorial PE applications.

Territorial initiative	PE method	Target territory
BlueAp - Bologna Local Urban Environment Adaptation Plan for a Resilient City	Participatory design	Bologna (Italy)
Citizen science Lab of Leiden University	Science Shops	Leiden (The Netherlands)
Sustainable Urban Mobility Plan Bremen 2025	Scenario planning	Bremen (Germany)
Brainport Smart District (BSD)	Living Labs	Helmond (The Netherlands)
Energy vision Murau	Guiding Visions technique for Agenda Setting	Murau region (Upper Austria)

PE application 1: Participatory design in Bologna

The BlueAp initiative (Bologna Local Urban Environment Adaptation Plan for a Resilient City) took place in Bologna (Italy) from 2012 to 2015, and aimed to address the challenges faced by the city in relation to climate change. The participatory design (co-design) of the responsible and adaptive strategy towards climate change entailed cooperation with both public and private stakeholders (Bono et al., 2015): public bodies, public and private companies, trade and consumer associations, university and schools, consortia, non-profit organizations, land reclamation authorities. The target stakeholders were engaged in the participatory development of the adaptation plan both upstream and midstream. The upstream engagement took place through various workshops, thematic sessions, round tables and surveys. During all these, the stakeholders and citizens of Bologna had the opportunity to express their own ethics concerns towards facing climate change and to actively participate in drafting the plan – thus providing input on what issues and territorial needs the BlueAp project should address. As for

midstream public engagement, particularly the citizens were engaged through the online app *PlayBlueAp*. This app –in the form of a social game– allowed citizens to share their own environmentally friendly activities under six main environmental themes and gain online ranking points. In this way, the public was able to collect and report data related to damage in the city due to specific climate phenomena, and provide input as to how the government could address future climate challenges. With reference to the results of the participatory design, citizens had the opportunity to acquire new knowledge about this specific scientific field, be inspired and take responsibility for the environmental activities in their city. The city of Bologna was likewise benefited, by being able to integrate citizens' concerns, ethics values and public input into the new local adaptation plan –thus providing an output genuinely considering the vulnerabilities and needs of the territory.

PE application 2: Science Shop in Leiden

The science shop of the University of Leiden is typically known as the Citizen Science Lab (CSLab), and brings together different stakeholder groups for co-creating new research (citizen science) projects. The specific territorial initiative of CSLab described in the present paper refers to addressing the territorial (and international) challenge of air pollution. The Leiden Science shop organised as its first activity in 2018 an international workshop engaging stakeholders from all over Europe: air pollution researchers, NGOs, citizen science experts, creative research experts, app developers and representatives from local/national/EU governments, among others. The aim of the science shop's activity was to initiate a co-creation process by combining the 'top-down' approach of projects initiated by scientists, and the 'bottom-up' activities initiated by society (Lorenz Center, 2018). Participants were engaged in an upstream way, since they participated in brainstorming discussions on the value of citizen science (e.g. through 'World Cafés') and in co-working sessions for developing project proposals (The Citizen Science Cost Action, 2018). These proposals aimed at indicating what kind of research and pilot projects could be initiated in the future for addressing air pollution. Therefore, non-experts in particular had the opportunity to: (a) provide their input on what kind of research directions and citizen science initiatives could be applied for enhancing air quality; (b) be enabled to take responsibility and control over their own environment. Regarding the outcomes of the CSLab's territorial initiative, these refer to co-creating promising plans for incubation and pilot projects at both territorial and international level, in order for them to afterwards turn into citizen science initiatives (that can be seen as midstream PE).

PE application 3: Scenario planning in Bremen

The Sustainable Urban Mobility Plan (SUMP) is a project implemented in the city of Bremen, Germany that encouraged the participation of regional actors and citizens, in order to form a new mobility plan for the city by 2025. The goal has been to develop a transportation system that will ameliorate the quality of the lives of Bremen's residents and tourists, and will further support sustainable mobility (e.g. cycling). The project engaged regional actors in several phases of the program through a variety of engagement methods. In terms of upstream engagement, SUMP Bremen 2025 organised citizen forums and public interest groups in order to define the goals of the project. Afterwards, additional participants were engaged for expressing their opinion on Bremen's opportunities and weaknesses – as displayed through a status analysis (SUMP Bremen 2025, 2021). The new participants were comprised of citizens who took part in regional committees or gave their input through an online portal. Then, the engaged actors were provided with 5 different "Test Scenarios'' (future scenarios). Each scenario presented an extreme case of a transport problem, accompanied by corresponding measures/possible solutions –these had been collected during the upstream engagement. In terms of

applying midstream engagement procedures, the participants evaluated all measures based on their potential effectiveness and in accordance to the goals of SUMP Bremen 2025. The engaged actors thus had the opportunity to indicate how (and through which specific directions) the target problem could be addressed After this step, a final "Target scenario" was developed and presented to the audience, who could now choose which measures should be featured in the final scenario. All actors were finally engaged in a downstream way, since they expressed their opinion towards the adoption of the final implementation plan and were further informed on three possible funding routes. They were asked to prioritize measures and decide which of them would be primarily or secondarily implemented. Overall, and through the scenario planning, SUMP Bremen 2025 was able to gain great input from its citizens who live and operate in the city; the engagement procedures did not focus on publicly presenting an almost-finished plan, but they provided to the public the opportunity to express their visions and integrate their concerns into every phase of the planning process. In other words, the plan ensured that transport solutions would continue to be formed according to the needs of Bremen.

PE application 4: Living Lab in Helmond

The Brainport Smart District (BSD) is a living lab in the city of Helmond in Eindhoven, the Netherlands, aiming to enhance smart development within the context of a community-building project. Its aims refer to forming an ameliorated living environment that enhances technological and sustainable innovation, and concurrently corresponds to the needs of its residents (Brainport Smart District, 2021a). At the same time, its ultimate goal can be described as co-creating a functional environment of an actual neighbourhood, developed as a safe and smart living space for its residents through eco-friendly strategies. The public engagement initiatives of BSD began in 2017, set to include different stakeholders within a span of 10 years (2018-2028) (Gebhardt, 2019). The project capitalised on a 'bottom-up' and 'top-down' approach, while the target stakeholders have so far been included in all stages of the project, particularly during the co-design and co-decision processes. Regarding the upstream engagement procedures applied, an initial workshop took place in 2017 (Syntegration), with the participation of different actors −regional and governmental actors, technology experts, STEM scientists, educational institutes, companies and inhabitants. The engaged actors produced twelve themes and provided their input toward the future directions of the living lab −since the final themes served as a baseline for the BSD program lines (Circular district, Participation, Social and Safe district, Healthy district, Digital district, Mobile district, District with Energy and District with water) (Brainport Smart District, 2021b). In the later phases of the BSD initiative, the aim is to engage different inhabitants from younger/older ages, with diverse backgrounds (e.g. lifestyle, income), but also businesses and employees who are going to provide feedback on how the conditions of their everyday life have been affected during the BSD program. The target stakeholders' engagement finally indicated that the upstream engagement was the most systematised, while there are also a few indications that downstream engagement will be applied in the remaining lifespan of the initiative. As for the overall outcomes of the engagement practices, the inhabitants have had the opportunity to influence and contribute to the design and development of the district, which will include approximately 1500 houses and a large-scale business park. Public participation has already assisted and will further assist BSD to create a smart neighbourhood exhibiting a high quality of living, potentially serving as an example for many areas internationally.

PE application 5: Guiding Visions technique for Agenda Setting in Murau

The Energy Vision Murau (EVM) is a project aiming to transform the area of Murau, upper Austria, into a self-sufficient region concerning energy use. The project started in 2002-2003 and formed an agenda ("Energievision Murau" 2015) with the input and contribution of different actors. A 'bottom-up' approach was adopted for both upstream and midstream engagement procedures, including interviews, workshops and events with a variety of actors –such as regional professionals related to renewable energies, installation, suppliers of energy, agricultural actors, different schools, politicians, people of administration and residents. In detail, the target actors were primarily engaged in an upstream way and worked together in order to define the objectives of Murau 2015. During this initial stage, the means for practically realising these objectives were also discussed. Proceeding a step further to the midstream engagement, a bigger group of actors (additional people joined) worked collaboratively in groups, in order to form strategies on how the project will proceed. Many of the actors were professionals related to energy and installation, and provided valuable input owing to their expertise. The participation of regional actors was also significant in the practical implementation stage of the Murau vision. Throughout the project, different actors formed additional networks and worked together towards common goals; for instance, political representatives collaborated with different energy companies. An example of such a cooperation refers to a company (member of Natur-Installateure), which started installing sustainable heating systems in new houses after working with a mayor and the project of Murau (Späth & Rohracher, 2010). Other project outcomes refer to the community strongly enhancing the EVM vision and processes; relevant achievements from regional stakeholders refer to one municipality being fuel-oil free and the overall region being self-sufficient in energy by 80%. Another example refers to the largest consumer of energy in the region (Stolzalpe Hospital) having increased the use of biomass for heating (it replaced over 1 million litres of oil). What should be overall noted is that the EVM engagement procedures enhanced the concept of 'energy systems of tomorrow', which is a tangible example of a simultaneous technical and social contribution –the so-called socio-technical integration in STI (de Saille, 2015).

DISCUSSION - CONCLUSIONS

Overall, territorial PE and related initiatives contribute to creating a new social contract between science, technology, innovation and society. Particularly due to the spatial/territorial context surrounding the initiatives, common territorial stakes can be addressed; in this way, the territorial STI processes become beneficial for the territory itself by being aligned to public concerns and ethics values. Likewise, the actors engaged (particularly the citizens) gain new knowledge and take responsibility for the development of their own territory. The present section reports some concluding observations on territorial PE –referring, among others, to most common engagement vantage points, to interlinkages with principles of equality and with the smart directionality approach – and on how it can be effectively applied for including ethical concerns to the transformational STI-related agendas of territories. These observations and concluding remarks can function as a basis or as an inspiration point for designing and implementing similar initiatives. More specifically, the PE initiatives examined cannot be entirely replicable, but the concluding arguments form a continuum; nothing is entirely replicable, but several aspects can be inspiring for PE potential implementers and for achieving the target transformation of their territory.

First of all, it has been noticed that the upstream engagement procedures are the most systematised ones for engaging the public and interested stakeholders at the territorial level. All territorial PE initiatives engaged the target stakeholders early in the process, so that experts and the public would co-define and co-design the goals of the initiative and its territorial orientation. In this way, public

concerns and ethical values would be considered prior to the actual implementation of STI territorial initiatives. This further comes in correspondence to general PE aspirations about early engagement before the target topics and final opinions become controversial, hardened, or polarised (Cobb & Gano, 2012). Subsequently, territorial PE initiatives that involve the target stakeholders in an upstream way can prove to be considerably beneficial, since the activities that will transform the territory will have considered in advance public ethics concerns. At the same time, this early engagement can create a sense of ownership to the engaged stakeholders, making them feel 'problem-owners' and genuinely mobilised in relation to the challenge addressed and initiating a series of territorial PE activities that will bring genuine impact.

Proceeding a step further, it is evident that fairness and equality principles play a pivotal role in the engagement procedures. All interested stakeholders would be engaged in the territorial PE initiatives, irrespective of their social status, cultural background, race, gender or even age at some instances. This indeed contributes to including an even wider range of perspectives and concerns, and to delivering outcomes that are ethically acceptable by a majority of people. Therefore, by enhancing fairness principles, the unfolding of the PE initiatives is context-dependent, in alignment to the regional context and characteristics (Wittrock & Forsberg, 2019), as well as in alignment to suggestions made towards a more effective implementation of RRI and its keys at various environments in the future (Gerber et al., 2020). It should be finally underlined that this emphasis on fairness, equality and on the inclusion of diverse individuals aids in creating new and powerful networks in the territory. Particularly due to the territorial scale, these networks of various actors and individuals have stronger ties, and contribute to a successful diffusion of knowledge about an ethically accountable science, technology and innovation.

As for the territorial challenges that PE initiatives tend to address, they can be listed among the Sustainable Development Goals (SDGs), often to the ones related to climate, environment and energy areas. As a follow-up to this observation, one can notice that territorial PE can successfully address the territory's' concerns and perspectives when combined to the smart directionality approach (Mazzucato, 2016). This approach suggests that knowledge production and exploitation should address societal goals and challenges. It further enhances the responsible and ethical use of research results for societal purposes. Consequently, the inclusion of public values and concerns into the STI processes can be the first step towards achieving the aforementioned responsible and ethical use – that can ultimately lead to sustainable R&I and STI investments in the target territories.

Along with these evidence-based remarks towards an effective territorial PE application, emphasis should also be placed on accompanying implications and limitations. Considerable obstacles might firstly be encountered when trying to engage the target audiences. Some stakeholders may be reluctant, sceptical or occasionally lack trust towards the local authorities. Particularly this lack of trust (or even mistrust) may very well stem from a rather old but long-standing perspective; this refers to the public participation being used as an additional argument for the legitimacy of pre-defined decisions by experts (Callon, Lascoumes, & Barthe, 2001). Concurrently, experts involved in the engagement procedures may unintentionally (or intentionally) reinforce the representation of the lay public as "ignorant" (Bucchi & Neressini, 2008). This scenario is even more prominent at territorial levels, where due to the smaller regional scale a few emblematic and leading figures do not allow space for counter-arguing their ideas. Such situations can particularly cause de-motivation and reluctance towards the active participation of the lay public and of their expressing of genuine concerns. A few final implications to be noted refer to the application of RRI as a whole at territorial levels. Each territory is different and any initiatives should be context-dependent, but as approaching to larger territorial scales (up to European ones), cultural differences add to problems of communication and

coordinated action (Fitjar, Benneworth, & Asheim, 2019). Then, referring to any territory and irrespective of its size, similar initiatives entailing expert-public participation cannot but fall under other ordinary political and democratic processes, which entail additional concerns and interests on behalf of actors involved –as similarly argued by Fitjar, Benneworth, & Asheim (2019).

Public engagement in R&I and STI processes can overall bring multiple benefits contributing to the accreditation of scientific and technological knowledge. In terms of application at territorial levels, this accreditation further adds to an ameliorated realisation of a common territorial stake, tailored to local ethical codes and values. Irrespective of the PE method(s) potentially selected for realisation of specific initiatives, PE implementers should bear in mind that expert and lay knowledge encounter each other and are not independent from each other. They are in need of hybrid forums –as Callon, Lascoumes, & Barthe (2001) call these places of interaction– so as to evolve, and it is highly possible that territories owing to their scale will be able to provide such forums in an effective way.

ACKNOWLEDGEMENTS

This work has been conducted within the framework of TeRRItoria (Territorial Responsible Research and Innovation Through the involvement of local R&I Actors) EU project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 824565.

KEYWORDS: Public Engagement (PE), Responsible Research and Innovation (RRI), Territorial, Regional, Ethics, Society

REFERENCES

Bono, L., Caranti, C., Fini, G., & Gueze, R. (2015). *Stakeholders Engagement Outcomes (Summary)*. Deliverable 1 - BlueAp - Bologna Local Urban Environment Adaptation Plan for a Resilient City. Retrieved from http://www.blueap.eu/site/wp-content/uploads/2013/09/B.1.1-SE-Outcomes-Summary.pdf

Brainport Smart District. (2021a, March 29). *Documents. Plan of Action*. Brainport Smart District. Retrieved from: https://brainportsmartdistrict.nl/en/organisation-business/documents/

Brainport Smart District. (2021b, March 29). *Programs*. Brainport Smart District. Retrieved from: https://brainportsmartdistrict.nl/en/organisation-business/programs/

Bucchi, M., & Neresini, F. (2008). Science and Public Participation. In E. J. Hackett, O. Amsterdamska, M. Lynch & J. Wajcman (Eds.). *The Handbook of Science and Technology Studies* (pp. 449-472). Cambridge, UK: MIT Press.

Caiati, G., & Mezzana, D. (2019). *Map of approaches, policies and tools for Territorial RRI*. Deliverable 3.3.-TeRRItoria project. Retrieved from http://territoriaproject.eu/wp-content/uploads/2021/04/TeRRItoria_D33_Map_of_approaches__policies_and_tools_for_Territorial_RRI.pdf

Callon, M., Lascoumes, P., & Barthe, Y. (2001). *Agir dans un Monde incertain: Essai sur la démocratie Technique*. Paris, France: Editions de Seuil.

Cobb, M. D., & Gano, G. (2012). Evaluating Structured Deliberations about Emerging Technologies: Post-Process Participant Evaluation. *International Journal of Emerging Technologies and Society*, 10, 96–110.

De Saille, S. (2015). Innovating innovation policy: the emergence of Responsible Research and Innovation. *Journal of Responsible Innovation,* 2(2), 152-168.

Dorstewitz, P. (2016). Imagining Social Transformations: Territory-making and the Project of Radical Pragmatism: Response to Review. *Contemporary Pragmatism*, 13(4), 361-381.

Elden, S. (2005). Missing the point: globalisation, deterritorialisation and the space of the world. *Transactions of the Institute of British Geographers*, 30(1), 8-19.

European Commission. (2020, February 10). *Public Engagement in Responsible Research and Innovation*. Horizon 2020. Retrieved from https://ec.europa.eu/programmes/horizon2020/en/h2020-section/public-engagement-responsible-research-and-innovation

European Commission. (2020, November 30). *Public Engagement*. Policy. Retrieved from: https://ec.europa.eu/research/swafs/index.cfm?pg=policy&lib=engagement

European Union, European Commission, & Directorate-General for Research and Innovation. (2013). *Options for strengthening responsible research and innovation*. Luxembourg: Publications Office of the European Union.

Fitjar, RD., Benneworth, P., & Asheim, BT. (2019). Towards regional responsible research and innovation? Integrating RRI and RIS3 in European innovation policy. *Science and Public Policy*, 46(5), 772-283.

Gebhardt, C. (2019). The Impact of Participatory Governance on Regional Development Pathways: Citizen-driven Smart, Green and Inclusive Urbanism in the Brainport Metropolitan Region. *Triple Helix Journal,* 6, 69-110.

Gerber, A., Forsberg, E.M., Shelley-Egan, C., Arias, R., Daimer, S., Dalton, G., Belén Cristóbal, A., Dreyer, M., Griessler, E., Lindner, R., Revuelta, G., Riccio, A. & Steinhaus, N. (2020). Joint declaration on mainstreaming RRI across Horizon Europe. *Journal of Responsible Innovation*, 7(3), 708-711.

Innerarity, D. (2013). Power and knowledge: The politics of the knowledge society. *European Journal of social theory,* 16(1), 3-16.

Laursen, K., Masciarelli, F., & Prencipe, A. (2012). Regions Matter: How Localized Social Capital Affects Innovation and External Knowledge Acquisition. *Organization Science*, 23(1), 177–193.

Lorentz Center. (2018). *Citizen Science Lab: Air Pollution*. Retrieved from: https://www.lorentzcenter.nl/citizen-science-lab-air-pollution.html

Marschalek, I. (2017). *Public engagement in responsible research and innovation*. [PhD dissertation, University of Vienna].

Mazzucato, M. (2016). From market fixing to market-creating: a new framework for innovation policy. *Industry and Innovation*, 23(2), 140-156.

Mejlgaard, N., Woolley, R., Carter, B., Bührer, S., Griessler, E., Jäger, A., Lindner, R., Bargmann Madsen, E., Maier, F., Meijer, I., Peter, V., Stilgoe, J., & Wuketich, M. (2018). Europe's plans for responsible science. *Science,* 361(6404), 761-762.

Mitcham, C. (2003). Co-responsibility for research integrity. *Science and engineering Ethics,* 9(2), 273-290.

MountEE. (2021, March 30). *Energy Vision Murau*. MountEE Sustainable Public Building. Retrieved from http://www.mountee.eu/good-practice/strategies/3-at/.

Mourey, T. (2015). *Bremen: SUMP monitoring and evaluation champion (Germany)*. Eltis. Retrieved from https://www.eltis.org/discover/case-studies/bremen-sump-monitoring-and-evaluation-champion-germany.

Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science and public policy*, 39(6), 751-760.

Paasi, A. (1998). Boundaries as social processes: Territoriality in the world of flows. *Geopolitics*, 3(1), 69-88.

Pellé, S., & Reber, B. (2015). Responsible innovation in the light of moral responsibility. *Journal on Chain and Net-work Science,* 15(2), 107-117.

Rip, A., & Shelley-Egan, C. (2010). Positions and Responsibilities in the 'Real' World of Nanotechnology. In R. Von Schomberg & S. Davies (Eds.). *Understanding public debate on nanotechnologies: options for framing public policies. A Report from the European Commission Services* (pp. 31-38). Brussels, Belgium: European Commission Services.

Sassen, S. (2013). When territory deborders territoriality. *Territory, politics, governance*, 1(1), 21-45.

Späth, P., & Rohracher, H. (2010). 'Energy regions': The transformative power of regional discourses on socio-technical futures. *Research Policy*, 39, 449–458.

Steinhaus, N. (Ed.). (2013). *Living Knowledge – Journal of Community-based research: Future Options for Responsible Research and Innovation*. Bonn, Germany: International Science Shop Contact Point.

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, 42(9), 1568-1580.

SUMP Bremen 2025. (2021, May 7). *English version: Verkehrsentwicklungsplan Bremen 2025 / SUMP Bremen 2025*. Bauumwelt Bremen. Retrieved from https://www.bauumwelt.bremen.de/mobilitaet/verkehrsentwicklungsplan-5586

Taebi, B., Correljé, A., Cuppen, E., Dignum, M., & Pesch, U. (2014). Responsible innovation as an endorsement of public values: the need for interdisciplinary research. *Journal of Responsible Innovation*, 1(1), 118–124.

The Citizen Science Cost Action. (2018). *Citizen Science Lab: Air Pollution - Meeting in Leiden*. Retrieved from: https://cs-eu.net/blog/citizen-science-lab-air-pollution-meeting-leiden

Wittrock, C., & Forsberg, E.M. (2019). *Handbook for Organisations Aimed at Strengthening Responsible Research and Innovation*. Deliverable 17.6 - RRI-Practice project. Retrieved from https://www.rri-practice.eu/wp-content/uploads/2019/06/RRI-Practice-Handbook-for-Organisations.pdf

COMPUTER ETHICS AND COMPUTER PROFESSIONALS

Norberto Patrignani, Iordanis Kavathatzopoulos

Politecnico of Torino (Italy), Uppsala University (Sweden)

norberto.patrignani@polito.it; iordanis.kavathatzopoulos@it.uu.se

Tech workers now want to know: what are we building this for?

Conger K., Metz C., 2018

INTRODUCTION

This paper investigates the intersection among the two domains: *computer ethics* (as an academic research field) and *computer professionals* (the professionals that work in real life organizations). The paper focuses around the main question: is the long history of *computer ethics* finally starting to "make a difference" in the real world of information and communication technologies (ICT) where computer professionals work?

COMPUTER ETHICS: A FIELD WITH A LONG HISTORY OF STUDY

Computer ethics as a research field has the same age of ICT: in the 1950s the "*social and ethical implications of computing*" were very clear to Norbert Wiener, one of the pioneers of the computer age (Wiener, 1950). Following Wiener, it is worth mentioning Joseph Weizenbaum, who in the 1970s, distinguished the problems that can be analysed and delegated to algorithms from the situations that cannot be "solved" by computational thinking, but require *judgment*, the "capacity to choose", a trait that makes people human (Weizenbaum, 1976). These recommendations appear visionary at a time when "artificial intelligence" applications are spreading and misused in many areas of society.

In the 1980s the term *computer ethics* was introduced by Maner: "*... a new field that studies ethical problems aggravated, transformed or created by computer technology*" (Maner, 1980). The critical absent from this definition are the *computer professionals*: ICT are just machines *designed by humans* that executes software code *developed by humans*. This limit in the initial approaches to *computer ethics* was still present in its formal definition proposed by Moor: "*there is a policy vacuum about how computer technology should be used*" (Moor, 1985). In this definition, technology in itself is not questioned, it is still considered as "neutral" and, again, *computer professionals* are in the shadow, even if they are the main agents as designers of complex ICT systems. Thanks to Deborah Johnson this connection became evident when she proposed to use the term "*socio-technical systems*" instead of "*computer systems*": according to her, technology is not neutral since technology and society co-shape each other: "*We have to keep stressing that engineering is a social activity*" (Johnson, 1985). Other contributions to the theoretical foundations of the academic field of *computer ethics* come from Bynum, who describes it as an applied ethics that relates to the: "*... identification and analysis of the impacts of information technology on such social and human values like health,...*" (Bynum, 1999), and Floridi who proposes an analogy between 'suffering in the biosphere' and 'entropy in the infosphere' and introduces the term *information ethics*: "*... what is good for an information entity and the infosphere in general? This is the ethical question asked by information ethics*" (Floridi, 1999).

This field has now produced a vast amount of research and is a well-established field with conferences and publications, but the main actors of this scenario, *computer professionals*, are still in the background.

COMPUTER ETHICS AND COMPUTER PROFESSIONALS

When *computer professionals* started to come to the foreground in the *computer ethics* debate? Probably the first researcher that focused on them was Donn Parker in the 1960s with its remarkable attention to people working with computers in real life and their relationship with ethics. He wrote: "*It seemed that when people entered the computer center they left their ethics at the door*" (Parker, 1968).

In the 1980s at *Xerox Palo Alto Research Center* (PARC) a discussion group, initiated by Severo Ornstein and Laura Gould, was formed by computer professionals concerned over the increasing role of digital technologies in war scenarios. This was the seed for the establishment of the *Computer Professionals for Social Responsibility* (CPSR), an organization dedicated to raising awareness among the profession and the public about the dangers of using computers in critical systems (Bruemmer, 1994). Unfortunately, CPSR as a membership organization dissolved in 2013 (Schuler, 2013).

In the 1990s, Donald Gotterbarn focused on "professionalism" and on the importance of: "*... the values that guide the day-to-day activities of computing professionals*" (Gotterbarn, 1991). He provided a fundamental contribution to the definition of a *Code of Ethics and Professional Conduct*, an important reference on "professional norms", released on July 2018 by the ACM (*Association for Computing Machinery*) (ACM, 2018).

In Europe this effort for "making a difference" in the real world of computer professional is becoming evident in the works of Simon Rogerson. He wrote: "*computer and information ethics are defined as integrating ICT and human values in such a way that ICT advances and protects human values*" (Rogerson, 2011). Ethics and ICT have to be strongly related and, by establishing in 1995 the *Centre for Computing and Social Responsibility* (CCSR) at De Montfort University, he gave a main contribution to the historical debate about the impact of strategic, managerial, and ethical issues of ICT inside real organizations.

Despite all these important efforts, recent events raise the question about the real impact of the *computer ethics* debate in the *computer professionals'* domain and inside business organizations.

EXAMPLES OF 'WORST PRACTICES'

One well known example of a "worst practice" is the *Volkswagen* "dieselgate". In September 2015 the EPA (*US Environmental Protection Agency*) communicated that the car manufacturer *Volkswagen* had installed a software for manipulating the data about car emissions with potentially dangerous consequences on human health. This kind of actions are strongly related to the attitude of *computer professionals*, and, according to Rogerson, put at risk their entire community's reputation. The importance of ethics in real contexts become evident: *computer professionals* must be aware about the risks of unethical practices (Rogerson, 2018).

Another well known example of software developers involved in a very controversial story is the "*Cambridge Analytica*": in March 2018 media around the world disclosed that the infamous *Facebook* app "*This is Your Digital Life*" distributed on the smartphones of 270,000 users of the social network was able to collect personal data of about 87 million of people by using the users' friends network. The

majority of these users were not aware about this data collection and the app developer gave all those data to *Cambridge Analytica* who used it for playing very controversial role in several political elections (Rosenberg et al., 2018). Without the skills of computer professionals all this data trick would not have been possible.

On October 2018 and on March 2019 two airplane crashes killed 346 people. Investigations discovered that the two airplanes, both Boeing 747 MAX, suffered a failure in the MCAS (*Maneuvering Characteristics Augmentation System*), a software for automated flight control activated by an erroneous indication from a sensor on the exterior of the airplane. Unfortunately, it was impossible for the pilots to regain the control (Laris, 2019). In this case, maybe the software plays just a minor role, nevertheless the entire system failed by keeping humans out of the loop.

In all these examples the question is: what is the role of *computer professionals* involved, even in very different roles, in these disasters? Were they aware of the consequences of the deployment of technologies coming from their work? A more general question arises: why people fail to adopt behaviours that demonstrate awareness about social and ethical issues in the ICT world?

HOW CAN WE RESTORE ETHICS IN ICT INDUSTRY?

The term 'pneumatophores' (inspired by botanical properties of some trees) is used for people who act as spirit bearers, who inspire others. Luckily in the ICT world, there are many 'pneumatophores', people who can be seen as exemplary by *computer professionals*. A study of moral exemplars in the computing profession showed that there are many dimensions involved in the moral excellence (Huff and Barnard, 2009). This paper is taking as a starting point just the words of Severo Ornstein, one of the leading computer scientists who worked at MIT's Lincoln Laboratory in 1955, Internet pioneer at Bolt Beranek & Newman in 1969, and founder of CPSR in 1983. In his words, from one side he is taking a strong position against the use of his knowledge and skills for military applications, when he says: "*I refused to work on classified projects*" (Bruemmer, 1994, p.4). He was a *computer professional* with a deep awareness about the context. From the other side, while looking around to his colleagues, he says: *"I also had for a long time been concerned that the whole profession seemed nerdish in the sense that it had its head down narrowly in what it was doing and playing games with it ... was not paying, seemingly, very much attention to what the social consequences of what it was doing ... people who seemed to be paying attention only to the thing itself and not very much to the context, and the context was clearly going to be the real world*" (Bruemmer, 1994, p.5). Sadly, in his words, he majority of *computer professionals* look without an awareness of the context.

Inspired by the 'pneumatophore' Ornstein, this paper proposes two main paths for restoring ethics in the ICT industry: one for people who are aware, and one for people who are not aware of the context.

COMPUTER PROFESSIONALS WHO ARE AWARE OF THE CONTEXT

In ICT industry, there are many people that feels being in conflict, while applying their competence and skills for developing morally controversial applications. At the same time, these *computer professionals*, across the technology industry ask questions that go beyond the technical details and functionalities. They are asking questions about purposes demanding greater insight into how their companies are deploying the results of their work (Conger and Metz, 2018). This often generates conflicts: more recently, Dr. Timnit Gebru, researcher in the Ethical Artificial Intelligence Team of Google, was fired by the company in December 2020. She was raising ethical concerns about the use of AI in language processing applications (Criddle, 2020). A dramatic conflict between ethical concerns

and real businesses. Here one could open a deep debate about the *business ethics* of many companies, but this would go beyond the scope of the paper.

Let's stay on the *computer professionals'* 'side'. It looks like there is an awareness about the social and ethical consequences of their work, but there are difficulties in "making a difference" in the real world. Maybe sometimes they feel alone or feel it is risky to act *personally*? Of course, especially for engineers working in large organizations, there is the well known approach called 'whistle-blowing'. When a person feels uncomfortable in collaborating with a project or when the employee recognizes unacceptable behaviours inside a company, then in some situations, the only possibility is 'to go outside' and let the public know about it. In particular when: the consequences for the public (or for the environment) are serious, all the warnings raised through the channels inside the organizations failed, there are impartial observers outside able to recognize the situation, and there is evidence that making the situation public will prevent the risks of serious consequences (De George, 1981). One of the best examples of 'whistle-blowing' in ICT world and engineering ethics is the 'Snowden case' in 2013 (Johnson, 2020).

In some situations, the single person feels being in real difficulties, feels alone, or the risk of losing the job is too high for the employee. So, what can a *computer professional* do in these cases? In many ICT companies, there are signs of *computer professionals* that start to organize their actions *collectively*. One example is the protest organized by Google workers in 2019 when they discovered that the company was collaborating in a military project (Hollister, 2019).

So, when *computer professionals* are aware about the context and feel being in conflict with their employers they can act *personally* or *collectively*. Here the role of international organizations like CPSR with its working groups like 'computers in the workplace' and 'technology and ethics' is fundamental. Nowadays, there is the ACM, with its Code of Ethics and related materials like 'The ACM Integrity Project: Promoting Ethics in the Profession', 'Ask an Ethicist", or 'Case studies' that can provide useful support (ACM, 2018).

COMPUTER PROFESSIONALS WHO ARE NOT AWARE OF THE CONTEXT

The question about the levelof awareness about the social and ethical implications of technology in the community of *computer professionals* of course requires further study and more systematic social research. This paper takes a simplified approach just to identify possible directions for addressing the issue. Back to the words of Ornstein ("*people ... paying attention only to the thing itself and not very much to the context*"), indeed in ICT world, there are a large number of computer professionals that like to concentrate just on the technical side.

Unfortunately, the view of technology as 'neutral' was present even in the approach on one of 'founders' of technology, John Von Neumann: "*I would prefer not to join the Board (of Bulletin of Atomic Scientists), since I have ... avoided all participation in public activities, which are not of a purely technical nature*" (Von Neumann, 1946). Indeed "*The Bulletin of the Atomic Scientists*", was established in 1945 to raise awareness about the risks for humanity related to technological advances (Boyer, 1985). On the other side, another 'founder', Norbert Wiener wrote: "*I do not expect to publish any future work of mine which may do damage in the hands of irresponsible militarists...*" (Wiener, 1947). A first suggestion for *computer professionals* could be: to read, for example, the letters of these two giants of the computer history (Heims, 1980).

Many ICT technical people are so much involved in their projects that could look a little 'naive', even if one can consider them as 'good guys'. This is the main reason for introducing *computer ethics* courses

even to undergraduate students in computer science: to raise awareness among 'good technicians' of the risks of underestimating the role of the context in ICT complex systems (Gotterbarn, 2015).

There are many approaches to teaching *computer ethics* in computer science courses, some of them are *concentrated* in one course, some are *distributed* among all the curriculum, but all of them try to enlarge the landscape of the stakeholders involved with ICT and reflect on their relationships (Patrignani, 2020; Karoff, 2019). All these efforts try to improve the *ethical competence* of future *computer professionals*.

Despite all these efforts at university level, there are issues about the general culture in ICT that need to be addressed: from one side the culture of 'innovation' and the other the culture of 'coding'.

About 'innovation' at any cost, often the mantra of ICT, indeed computer professionals need a kind of 'antidote' to the 'motto' of the high-tech industry "*disrupt first, ask questions later*". Later could be *too late*, and the capability to 'anticipate' is becoming a requirement in the engineering field (Pasquale, 2020).

About 'coding', this 'buzzword' is becoming popular in the schools at any age. Indeed, it is true that we need to educate children, even at elementary schools providing them with a good level of 'digital literacy' in order to avoid the risk of becoming just 'digital consumers'. But for the future generations, we need also to introduce a 'wise use of technology' even at primary level schools.

According to Rogerson, it is time to go beyond *computer ethics* just for computer scientists, a renewed *digital ethics* is needed for everyone since childhood: "*... In the digital age it is people ... who make digital technology... digital ethics education in the post millennial era is best started from early childhood*" (Rogerson, 2021).

If we want to prepare future generations for a true 'digital citizenship', then introducing scientific method, problem solving, computational thinking, and coding is necessary, but not enough. We need to introduce also digital identity management skills, a good balance between online and offline time, a knowledge about the use of technology for the public interest (like 'ethical hackers'), and a Slow Tech approach: a good (socially desirable), clean (environmentally sustainable) and fair (ethically acceptable) ICT (Patrignani, 2020). There are several interesting experiments in this direction, starting from 'teaching the teachers' of elementary schools like the project "*From coding to digital wisdom*" (Loccioni, 2021). Maybe the time has come for *computer ethics* or *digital ethics* for all ages?

CONCLUSIONS AND NEXT FUTURE

The role of ICT in this century is becoming pervasive, with seven out of the ten greatest companies in the world that are the "titans of the Web". ICT has a critical impact on society and the environment. They are now the main challenges of the Anthropocene.

Computer professionals represent the core node of the ICT stakeholders' network, they are at the centre of fundamental design choices, but usually operate inside large organization: how can their "ethical competence" be balanced with the power of business companies, how can they cope with the power of the organizations in which they operate? Can they act just individually? Maybe it is time to re-establish organizations like *Computer Professionals for Social Responsibility*? How can they develop their "moral autonomy" in social situations where both *thinking* and *action* are involved? How can we educate *computer professionals* to acquire the necessary ethical skills? (Kavathatzopoulos, 1988, 2003; Patrignani, 2020). Maybe it is time to introduce mandatory *computer ethics* courses for undergraduate students? For *all* education levels?

KEYWORDS: Computer Ethics, Computer Professionals, Volkswagen, Snowden.

REFERENCES

ACM (2018). *ACM Code of Ethics and Professional Conduct* (last update). Retrieved from https://www.acm.org/code-of-ethics

Boyer, P.S. (1985). *By the bomb's early light: American thought and culture at the dawn of the atomic age*. Pantheon.

Bruemmer, B.H. (1994). *An Interview with Severo Ornstein and Laura Gould*. Charles Babbage Institute, Center for the History of Information Processing, University of Minnesota.

Bynum, T.W. (1999). Keynote at AICE99, The Foundation of Computer Ethics. *Computer and Society*, June 2000.

Conger, K., Metz, C. (2018, October 7). Tech workers now want to know: what are we building this for? *The New York Times*.

Criddle, C. (2020, December 8). Thousands more back Dr Timnit Gebru over Google 'sacking'. *BBC News*. Retrieved from https://www.bbc.com/news/technology-55164324

De George, R. (1981). Ethical Responsibilities of Engineers in Large Organizations: The Pinto Case. *Business and Professional Ethics Journal*, 1, no.1.

Floridi L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology*, 1: 37–56, Kluwer Academic Publisher.

Gotterbarn, D. (1991). A "Capstone" Course in Computer Ethics, in (eds.) Bynum, T.W., Maner, W., Fodor, J.L. (1991), *Teaching Computer Ethics*, Research Center on Computing and Society, S. Conn. State Univ.

Gotterbarn, D. (2015). *What is Applied Ethics Good For?* CEPIS Ethics Conference 2015.

Heims, S.J. (1980), *John Von Neumann and Norbert Wiener, from mathematics to the technologies of life and death*, MIT Press.

Hollister, S. (2019, November 25). Google is accused of union busting after firing four employees. *The Verge*.

Huff, C. Barnard, L. (2009). Good computing: moral exemplars in the computing profession. *IEEE Technology and Society*, 28(3).

Johnson, D.G. (1985). *Computer Ethics*. Prentice-Hall.

Johnson, D.G. (2020). *Engineering Ethics: Contemporary and Enduring Debates*. Yale University Press.

Karoff, P. (2019, January). Embedding ethics in computer science curriculum, *Harvard Gazette*.

Kavathatzopoulos, I. (1988). *Instruction and the development of moral judgement*. Uppsala Universitet, Almqvist & Wiksell International, Stockholm, Sweden.

Kavathatzopoulos, I. (2003). The Use of Information and Communication Technology in the Training for Ethical Competence in Business. *Journal of Business Ethics*, 48: 43-51.

Laris, M. (2019, June 20). Changes to flawed Boeing 737 Max were kept from pilots, DeFazio says. *The Washington Post*.

Loccioni (2021). *Training teachers. From coding to digital wisdom.* https://www.loccioni.com/en/waves/dal-coding-alla-saggezza-digitale/

Maner, W. (1980). *Starter Kit in Computer Ethics*. Helvetia Press and the National Information and Resource Center for Teaching Philosophy.

Moor, J. (1985). What Is Computer Ethics? *Metaphilosophy*, 16(4): 266-75.

Parker, D. (1968). Rules of ethics in information processing, *Communications of the ACM*, March 1968 (Vol. 11, No. 3).

Pasquale, F. (2020). *New Laws of Robotics. Defending Human Expertise in the Age of AI*. Harvard University Press.

Patrignani, N. (2020). *Teaching computer ethics. Steps towards Slow Tech, a good, clean, and fair ICT*. Uppsala University, Acta Universitatis Upsaliensis.

Rogerson, S. (2011). Ethics and ICT, in (eds.) Galliers R.D., Currie W., *The Oxford Handbook of Management Information Systems. Critical Perspectives and New Directions*, "Part IV: Rethinking MIS Practice in a Broader Context", Oxford University Press.

Rogerson, S. (2018). Ethics omission increases gases emission: A look in the rearview mirror at Volkswagen software engineering. *Communications of the ACM*, Vol. 61 No 3, March, pp30-32.

Rogerson, S. (2021). Rebooting ethics education in the digital age. *Academia Letters*, Article 146. https://doi.org/10.20935/AL146.

Rosenberg, M, Confessore, N., Cadwalladr, C. (2018, 17 March). How Trump Consultants Exploited the Facebook Data of Millions. *The New York Times*.

Schuler, D. (2013). *Dissolution and Gary Chapman, Winner of CPSR's Norbert Wiener Award*. Public Sphere Project, May 2013.

Von Neumann, J. (1946). *John Von Neumann to Norman Cousins* (Library of Congress archives). May 22, 1946. In Heims, S.J. (1980), *John Von Neumann and Norbert Wiener, from mathematics to the technologies of life and death*, MIT Press.

Weizenbaum, J. (1976). *Computer power and human reason: from judgment to calculation*. Freeman.

Wiener, N. (1947, January). A Scientist Rebels. *Atlantic Monthly*.

Wiener, N. (1950). *The human use of human beings*. The Riverside Press.

REASONS FOR RESISTING THE ACCEPTANCE OF HYPERNUDGES

Yukari Yamazaki

Seikei University (Japan)

yyamazak@bus.seikei.ac.jp

ABSTRACT

As the development and prevalence of artificial intelligence and machine learning (AL/ML) have progressed at a surprising speed, the acceptance of AI artefacts has drawn the attention of scholars. While early arguments on the technology acceptance model and human–computer (or robot) interactions (HCI) have discussed the factors promoting acceptance and utilisation of new technologies, there has been another stream of thought about the factors evoking avoidance or resistance of systems.

In this study, we investigate whether several AI-driven hypernudges are acceptable to people as well as the reasons users resist several AI services. A total of 1,211 participants were asked to answer whether they agree or disagree with 16 types of hypernudges of various categories, depth, and types. In addition, they were asked to choose eight reasons for their resistance to AI use. The results showed that 4 out of 16 hypernudges were rejected, but 6 hypernudges were significantly more acceptable to more than half of the participants. There are three significant reasons why participants resisted AI use: using personal information without the consent of the user, AI suggestions do not always coincide with the user's feelings at that time, and unpredictability of the consequences if users accept AI support. These results indicate that people tend to neither resist hypernudges driven-by AI nor are concerned about AI support, except for privacy and consistency with their own feelings. Additionally, we will discuss the key shortcomings of hypernudges thus far and the direction of future research.

INTRODUCTION

Over the past several decades, considering the development, penetration and utilisation of AI/ML in various scenarios has been common in academic and practical fields. While the efficiency, effectiveness, and convenience of AI artefacts have been discussed or broadcasted on multiple media, the acceptability of AI supports that are on behalf of human autonomy has as yet been scarcely dealt with. Because AI has its own peculiar features such as autonomy, the black box of its mechanism (algorithm), and the unpredictability of its consequences, it deserves to be refused or avoided by people. Moreover, despite the facts that several serious legal and ethical matters, such as lethal autonomous weapons, might be brought on by AI artefacts, whether people hesitate to use it has not been revealed. Meanwhile, it is concerning that people might welcome AI support without deeply considering the consequences. In this study, we investigate the reasons people choose to resist the support of AI artefacts.

CURRENT STUDY

There are approximately two streams of discussions on the avoidance or resistance to new technologies or systems. The first has been issued in the fields of human–computer (or human-robot) interaction (HCI), and the second has been argued in organisational contexts.

The research on the former has shed light on the traits of autonomous technologies and has determined that factors such as a system's transparency (Kim and Hinds, 2006), controllability (Jameson and Schwarzkopf, 2002), trust (Lewandowsky et al., 2000), prior interaction experience (Kirchbuchner et al., 2015), physical contact (Evers et al., 2010), or shared driving goals (Verberneet al., 2012) play a role in how people perceive and resist autonomous technologies.

The research in HCI, as the first issues on resistance of systems, indicates that autonomous systems make individuals feel controlled physically and psychologically (Stein et al., 2019), restrict their freedom (Kang, 2009), and take away jobs meant for humans (Waytz and Norton, 2014). In the examination of medical patients, participants were reluctant to utilise healthcare delivered by AI. This is because an automated adviser evokes a mismatch between two fundamental beliefs, wherein, patients view themselves as unique and different from others; contrarily, machines try to standardise and treat every case in the same way (Logoni et al., 2019). Another stream of *dehumanisation* research has examined the significance of human uniqueness for people's self-esteem (e.g., Ferrari, et al., 2016; Vaes et al., 2003). An advice-giving robot that provided highly threatening advice messages to human autonomy invoked feelings of anger and negative thoughts compared to low threat-to-autonomy advice messages (Roubroeks et al., 2011). In particular, autonomous robots may be perceived as lacking predictability, which would be an additional hazard for human safety. Individuals tend to show negative attitudes towards emotional interactions with robots (Zlotowsk, et al., 2017). Furthermore, individuals are more likely to attribute responsibility to the robot when they perceive the robot to be autonomous (Kim and Hinds, 2006).

The research on the organizational contexts, as the second issues on resistance of systems, has mainly proposed guidelines toward top management, have treated acceptance and resistance as a dichotomy (Lapointe and Beaudry, 2014) and applied the technology acceptance model. Markus (1983) defines resistance to IT as ‘behaviours intended to prevent the implementation or use of a system or to prevent system designers from achieving their objectives’ (p. 433). Even though, interest in the phenomenon of resistance to systems appears to be growing, resistance has received relatively little attention when compared to acceptance. Therefore, similar to the technology acceptance model that considers the factors that promote the use of information technology, the degree of resistance or reluctance to utilise information technology is also a function of the factors for experience with similar technology (Martinko et al., 1996), systems being implemented (Markus, 1983), the context (e.g., individual or group) of its use (Joshi, 1991), stress or pressure for misuse (Marakas and Hornik, 1996), and so forth. Several studies have treated IT resistance as a behaviour to express use (Kane and Labianca, 2011), other studies have tried to conceptualise the resistance of cognitive (Bhattacherjee and Hikmet, 2007), psychological (Lorenzi and Riley, 2000), attitudinal (Robey, 1979), and affective (Selander and Henfridsson, 2012) forces but not behavioural. Overall, since resistance represents a normal psychological reaction when a person perceives the consequences of system implementation or new technologies as negative (Lorenzi and Riley, 2000) or a threat (Kim and Kankanhalli, 2009), the guidelines for overcoming resistance would be constructed as models (e.g., Lapointe and Rivard, 2005).

The AI artefacts that this study has focussed on are not only autonomous systems, but they also possess other peculiar features that make people hesitant to use them. Five factors have been identified as other concerns that might emanate for the reluctance to use AI-driven artefacts. The first is incomprehension, which comes from the black box nature of AI/ML computational mechanisms, even when the algorithm has opened to the public (Pasquale, 2015). Currently, AI/ML has reached capabilities that go beyond a level that a human being can fully comprehend without a detailed understanding of the underlying mechanisms (Arndras et al., 2018). The second is the inconsistency between the recommendations and the user’s feelings at the time because of the dynamic features of systems (Lanzing, 2018). Users are unsure whether these are acceptable or accurate because

recommendations based on purchase histories or online behavioural advertising, from multiple sources, may vary rapidly and dynamically. Additionally, the preferences of users do not always change symmetrically or simultaneously with data presented by apps based on AI/ML.

The third factor is the lack of informed consent for the use of data resources that AI/ML utilises. Emerging technologies and online hypernetworks have spread throughout the world, and widespread and ubiquitous collection of data has become commonplace. Consequently, several individuals are unaware of where and how their data are gathered and used (Lanzing, 2018). It is not surprising that, when users use novel technologies, without hesitation, they rely on informational boundaries and norms that predate the big data era (Patterson, 2013). Data, however, may be used for other (occasionally harmful) purposes in the future, which cannot currently be foreseen (Van der Sloot, 2017; Agarwal et al., 2013). The fourth factor concerns decreasing motivation (Maier and Seligman, 1976), moral conflict (Wertenbroch et al., 2008), and lack of responsibility (Greene and Cohen, 2004), which depends on the autonomy of the systems. In addition, there are concerns about indulging in hedonism (Chen and Sengupta, 2014) and confusing the user whether the decision making is as per their own will when delegating autonomy to systems and accepting recommendations from others; in other words, heteronomy (Raz, 1986), is pointed out as the fifth factor.

DICISIONAL INTERVENTIONS USING AI ARTEFACTS

Linkages between nudges (nudging) in behavioural science and hyper technologies such as self-tracking, utilising location information, recommendation systems using behavioural history, and AI have been propounded as a neologistic word, Hypernudges. On the one hand, steering people in a better direction using the tremendous power of big data and algorithms seems to be a convenient, effective, and wonderful service. On the other hand, hypernudges might cause several unintended side effects; as is said that good medicine tastes bitter. Indeed, quite a few studies have argued about the negative effects and concerns of hypernudges. Examples of concerns for hypernudges include the legitimacy and legal implications of these techniques (Yeung, 2017), invasion of privacy (Lanzing, 2018), personalisation (Peer et al., 2019; Mills, 2019), and ethical and philosophical matters in hypernudges (Sætra, 2019).

Both nudges and hypernudges are basically gentle and mild, neither is obtrusive or disturbing, and they promote better direction. However, it is necessary to pay attention to two aspects. The first concerns the depth of interventions. Several prior researchers have found that people tend to disagree with deeper (excessive) interventions, such as mandatory choice architecture, than shallower ones (Yamazaki, 2020) as well as unconscious (overt or transparent) interventions where people are unaware of being supported than conscious ones (Felsen et al., 2013; Jung and Meller, 2016).

In the same vein, transparency of the nudge is required for the acceptance of nudges as well as for the avoidance of manipulation. A non-transparent nudge is defined as a nudge working in a way that the person in the situation cannot reconstruct either the intention or the means by which a behavioural change is pursued (Hansen and Jesperse, 2013). In addition to the transparency of nudge, the nudge related to the dual process theory and the two types of ways of human thinking and behaving will also affect the acceptance of nudges (Hansen and Jesperse, 2013). According to this notion, nudges can be divided into four types, based on the degree of transparency and the two types of systems (automatic versus reflective systems). Furthermore, the nudge that enacts System 1 (automatic system) with an untransparent method, such as the opt-in/opt-out for organ donation, must be manipulated.

Second, there are several features that make hypernudges different from nudges. Because hypernudges utilise big data, such as individual purchase history, they are too personalised. In addition,

such information is dramatically updated every moment, and feedback from hypernudges is dynamic, synchronous, and flexible. The other features are in line with the traits of AI, such as autonomy, unpredictability, and complexity of the mechanism because hypernudges might utilise AI artefacts.

RESEARCH QUESTIONS

To reveal the reasons that people hesitate to be supported by hypernudges through AI artefacts, in this study, we developed on prior survey studies that examined the acceptance of traditional nudges (Thaler and Sunstein, 2008). One unique nudge on information security and fifteen nudges on health, ecology, and donation from Sunsiten et al. (2018) are applied to and introduced to investigate people's resistance to AI utilisation. Prior studies that surveyed citizens of various countries on several types of nudging questions have found that a consensus on nudges was obtained, even though there were several controversial, manipulative or objectionably paternalistic objections (e.g., Sunstein, 2016; Sunsiten et al., 2018). According to prior studies on the resistance of new technologies in the fields of HCI and in organisational contexts, in this study we selected eight reasons, considering the features of AI artefacts and hypernudges. Thus, the following hypotheses were formulated:

H1a: Perceived opacity and incomprehensiveness toward AI artefacts are chosen as reasons for resisting their use.

H1b Those who choose 'opacity' as a reason for resistance tend to resist using AI artefacts.

H2a Perceived mismatch between the proposals by AI artefacts and their own feelings at that moment is chosen as a reason for resisting their use.

H2b Those who choose 'mismatch' as a reason for resistance tend to resist using AI artefacts.

H3a: Perceived unpredictability of the consequences is chosen as a reason for resisting the use of AI artefacts.

H3b Those who choose 'unpredictability' as a reason for resistance tend to resist using AI artefacts.

H4a Perceived ambiguity as to whether or not people have consented to use personal information for AI artefacts is chosen as a reason for resisting their use.

H4b Those who choose 'ambiguity of their consent' as a resistance reason tend to resist using AI artefacts.

H5a Demotivation of their own behaviour because of suggestions from AI artefacts is chosen as a reason for resisting their use.

H5b Those who choose 'demotivation' as a resistance reason tend to resist using AI artefacts.

H6a Indulging themselves because of recommendations from AI artefacts is chosen as a reason for resisting their use.

H6b Those who choose 'indulging' as a reason for resistance tend to resist using AI artefacts.

H7a Wondering whether it coincides with their own will because of recommendations from AI artefacts, is chosen as a reason for resisting their use.

H7b Those who choose 'wondering own will' as a resistance reason tend to resist using AI artefacts.

H8a Indifference to AI artefacts is chosen as a reason for resisting their use.

H8b Those who choose 'indifferent' as a resistance reason tend to resist using AI artefacts.

H9 The depth and consciousness of interventions influences the resistance of hypernudges.

PARTICIPANTS

A total of 1,211 participants, including 661 men and 550 women, were recruited through a Japanese consulting firm and a Japanese university. All participants are Japan-based, in the age range of 13-87 years (*M*=36.64, *SD*=18.2, under 20=332 (27.42%), 20s=212 (17.51%), 30s=143 (11.81), 40s=180 (14.86%), 50s=165 (13.63%), 60s=120 (9.91%), 70s=51 (4.21%), and 80s=8 (0.66%)). They received vignettes with two series of questions. The first question asked participants if they resist the following concrete interventions related to their health, the environment, or charity to enable participants to realise the support of AI artefacts at the beginning. Because there were several questions, around half of the participants (615 people) answered the first eight questions. The latter half (595 people) answered the remaining eight questions to mitigate participants' burden and avoid crude answers. The second question asked participants to choose the reasons for their refusal of AI support.

To enlarge on prior studies that considered the acceptance of nudges, in this study, we applied the interventions in previous research and added minor changes, namely, AI-driven hypernudges. In addition, one unique question about mobile information security was appended. A summary of the hypernudges is shown in Table 1. In this study, we considered the depth and types of interventions according to prior studies. The depth that suggests the degree of intervention is divided into five levels, in descending order: (5) Forced (not nudge), (4) mandatory choice architecture, (3) mandatory default, (2) mandatory information disclosure, and (1) campaign, (Sunstein et al., 2018). Consciousness is also categorised into two types according to Jung and Mellers (2016). Interventions that seem relatively difficult to notice or recognise that is, interventions on the default setting (No. 3, 9, 11, 13, and 16), availability (No. 5), and subliminal advertisement (No. 8), are categorised as unconscious. The participants answered whether they rejected each intervention with 'agree' or 'disagree.'

Table 1. Summary of Hypernudges.

	Contents of hypernudges	categories	depth*1	types*2
1	Online food shopping sites show calorie labels based on your purchase history utilising AI.	health	(2)	C
2	Online food shopping sites show bad effects of food on your health considering personal health conditions.	health	(2)	C
3	Enrolling green energy suppliers automatically that present more expensive energy according to users' income, possible to opt out.	ecological activity	(3)	S
4	Asking to be an organ donor according to your personal information in obtaining driver's licence.	Social security	(2)	C
5	Placing healthy foods based on your health condition analysed by AI at prominent visible positions in online food shopping sites.	health	(4)	C
6	An education campaign to reduce distracted driving based on your past driving history.	safety	(1)	C
7	An education campaign for promoting healthier choices for parents based on their children's health history to reduce obesity.	health	(1)	C
8	Providing prohibited subliminal advertisements on online movie sites to discourage your smoking and overeating based on your behavioural history.	health	(5)	S
9	Charging a specific amount with offset opt out option for carbon emission according to your boarding frequency.	ecological activity	(3)	S
10	Labelling unhealthy food based on your purchase history to enable you to notice that it is harmful.	health	(2)	C
11	Asking to donate to the Red Cross fund automatically according to individual income, possible to opt out.	social security	(3)	S

12	Requiring video sites to provide public education messages to discourage smoking and overeating based on personal and behavioural information.	health	(2)	C
13	Requiring large electricity providers to make people enrol in green energy suppliers automatically according to users' income, possible to opt out.	ecological activity	(3)	S
14	Keeping top page in online shopping sites free of unhealthy foods based on individual health condition to halt obesity.	health	(2)	C
15	Websites of public institutions show the message to have one meat-free day per week considering personal medical histories.	health	(4)	C
16	Installing security software according to personal usage history automatically to avoid viruses and hackers, possible to opt out.	information security	(3)	S

*1: It shows from the deepest order from (5).
*2: S suggests subconscious types and C suggests conscious types.

Participants were also presented the eight reasons and concrete suggestions (Table 2) to resist the support by AI artefacts and asked why they chose the reasons.

Table 2. Reasons to resist AI use.

	Reasons	Suggestions
1	Opacity and incomprehension	I neither know nor understand the AI mechanism itself or the reasons AI suggests it to me.
2	Mismatch	The suggestions by AI do not always coincide with my feelings at that moment.
3	Unpredictability	I cannot predict the consequences in case I follow the suggestions from AI. (AI suggestions may mismatch my assumption or hope.)
4	Without consent	AI might use my behaviour history and personal information without my consent.
5	Demotivation	Because the choice was made by AI, I lose both motivation and interest.
6	Corruption or Indulging themselves	Because of suggestions by AI, I may accept it though I think it is bad for me.
7	Wondering if it is my own will	Once relying on AI suggestions, I was confused as to whether it was my intention.
8	No interests	I do not have interests in AI artefacts (therefore I never accept it).

RESULTS

Overall resistance reasons and rejection of hypernudges

As shown in Table 3, first of all, it appears that the participants in this study are not devoid of an interest in AI because the selection rate of the reason 'no interest' is very low (16.68%, $\chi 2(1) = 537.778$, $p<0.01$). This result does not support H8a. Only two reasons, 'mismatch' (65.73%, $\chi^2(1) = 119.869$, $p<0.01$) and 'without consent' (56.56%, $\chi^2(1) = 20.876$, $p<0.01$) were chosen by more than half of *the participants. The reason that was least chosen is 'indulging yourself' (16.43%,* $\chi 2(1) = 545.804$, $p<0.01$), and the next was 'demotivation' (20.86%, $\chi^2(1) = 410.425$, $p<0.01$). 'Opacity' (32.29%, $\chi 2(1) = 151.974$, $p<0.01$) was also chosen relatively fewer times among all the reasons. In addition, 'unpredictability' ($\chi 2(1) = 3.489$, $p<n.f.$) has no significant effect as a refusal reason. Therefore, only H2a and H4a are supported; H1a, H3a, and H5a to H7a are not supported.

Table 3. Descriptive statistics and significance testing for choosing refusal reasons for AI.

Reasons	Number of choosing (*n*=1,211)	Percentage (%)	$\chi^2(1)$
(1)	391	32.29	151.974**
(2)	796	65.73	119.869**
(3)	573	47.32	3.489*
(4)	685	56.56	20.876**
(5)	253	20.89	410.425**
(6)	199	16.43	545.804**
(7)	489	40.38	44.83**
(8)	202	16.68	537.778**

*: $p < .05$
**: $p < .001$

Table 4 shows the results for rejecting hypernudges. No. 15 (requiring meat-free day) was the highest (70.57%), followed by No. 8 (subliminal advertisement) (66.61%). The former hypernudge was widely rejected because it intervenes with individuals' food choices, even though it was not deep and conscious. The latter intervention was also rejected due to forced instead of nudged. Furthermore, No. 14 (free of unhealthy food on top page) (60.00%) and No. 5 (placing healthy food at prominent positions) (54.87%) were also relatively higher than others. Both hypernudges are health categories, have shallower depth levels (2), and conscious types. It could be assumed that people are bothered by continuous recommendations during shopping and reject it. Over half of the participants disagreed with these four interventions. These results do not support H9, which suggests the effects of depth and types of hypernudges did not recognized.

In contrast, No. 16 (mobile security) (22.28%), No. 7 (reducing childhood obesity) (34.73%), and No. 10 (unhealthy food labelling) (35.28%) had low rejection rates, which are relatively acceptable to people. Six hypernudges were significantly accepted for over half of the participants; these included No. 2, No. 3, and No. 13. Neither deterministic nor common tendencies of category, depth, and types of hypernudges between higher and lower rejections were confirmed. The former three hypernudges might be considered as beneficial, familiar, and acceptable support by people.

No. 1, 4, 6, 9, 11, and 12 were not significantly different between ‘agree’ and ‘disagree’ selections.

Table 4. Descriptive statistics and significance testing for the rejection of 16 interventions (hypernudges).

No.	Number (1~8: *n*=596)	Percentage (%)	χ^2	No.	Number (9~16: *n*=615)	Percentage (%)	χ^2
1	290	48.66	0.43	9	310	50.41	0.08
2	269	45.13	5.644*	10	217	35.28	53.27**
3	244	40.94	19.57**	11	305	49.59	0.015
4	281	47.15	1.94	12	301	48.94	0.275
5	327	54.87	5.644*	13	267	43.41	10.668*
6	315	52.85	1.94	14	369	60.00	24.6**
7	207	34.73	55.577**	15	434	70.57	104.08**
8	397	66.61	65.779**	16	137	22.28	189.075**

*: $p < .05$
**: $p < .001$

Interactions Between Resistance Reasons and Rejection of Hypernudges

Next, the interactions between the resistance reasons and rejection of hypernudges were analysed. The result shows that there were three overall reasons for resistance: mismatches in suggestions from

AI, the unpredictability of AI, and lack of consent (Table 5). That is to say, the other five reasons: opacity, demotivation, indulging, inconsistent with own will, and no interest, were not chosen as reasons for rejecting AI by our subjects. These three reasons have significant effects on the rejection of hypernudges. Among them, the subjects who chose 'without consent' as a resistance reason rejected 10 out of 16 hypernudges (No. 1, 2, 4, ,5, 8, 9, 11, 12, 14, and 15). Those who chose 'mismatch' as a resistance reason rejected 7 out of 16 hypernudges (No. 1, 5, 8, 11, 12, 14, and 15). Those who chose 'unpredictability of the consequence' rejected 2 out of 16 hypernudges (No. 5 and 8). The hypernudges that were declined by people for these three reasons are No. 5 and No.8, which have a high rejection rate overall. No. 1, 2, 11, 12, 14, and 15 were rejected by the participants who chose 'mismatch' and 'without consent' as the resistance reasons. Significant differences between choosing reject reasons and disagreement of hypernudges are not confirmed among the other reasons. These results partially support H2b, H3b, and H4b.

Table 5. The results of Interactions between Three significant Reasons for Rejecting Hypernudges.

Reasons / Nudge		Mismatch			Unpredictability			Consent		
		Chosen	Not	$\chi^2(1)$	Chosen	Not	$\chi^2(1)$	Chosen	Not	$\chi^2(1)$
1	reject	210	80	7.188**	150	140	4.404	197	96	24.638**
	accept	190	116		132	174		143	163	
2	reject	196	73	7.34**	136	133	2.067	179	90	19.950**
	accept	204	123		146	181		158	169	
3	reject	170	74	1.225	125	119	2.539	151	93	4.797**
	accept	230	122		157	195		186	166	
4	reject	197	84	2.157	149	132	6.953**	173	108	5.458*
	accept	203	112		133	182		164	151	
5	reject	231	96	4.086**	170	157	6.345*	206	121	12.279**
	accept	169	100		112	157		131	138	
6	reject	216	99	0.643	151	164	0.103	182	133	0.414
	accept	184	97		131	150		155	126	
7	reject	149	58	3.403	117	90	10.783**	135	72	9.711**
	accept	251	138		165	224		202	187	
8	reject	280	117	6.282**	206	191	9.978**	249	148	18.462**
	accept	120	79		76	123		88	111	
9	reject	206	105	0.937	163	148	6.551**	199	112	14.031**
	accept	190	114		128	176		149	155	
10	reject	148	69	2.126	105	112	0.154	141	76	9.612**
	accept	248	150		186	212		207	191	
11	reject	210	95	4.769**	168	138	14.056**	190	116	7.516**
	accept	186	123		123	186		158	151	
12	reject	212	89	9.385**	150	151	1.498	186	115	6.511**
	accept	184	130		141	173		162	152	
13	reject	183	84	3.543	142	125	6.514**	159	108	1.689
	accept	213	135		149	199		189	159	
14	reject	263	106	19.063**	181	188	1.113	229	140	11.253**
	accept	133	113		110	136		119	127	
15	reject	310	124	31.859**	215	219	2.921	259	175	5.739
	accept	86	95		76	105		89	92	
16	reject	91	46	0.318	80	57	8.676**	89	48	5.036
	accept	305	173		211	267		259	219	

*: $p < .05$

**: $p < .001$

The coloured cells show significant differences between the resistance reason chosen and rejection of hypernudges.

As a representative feature of the results, the hypernudges that have low rejection rates do not have a significant interaction with the resistance reason chosen. On the one hand, it is revealed that people are concerned about the protection of their privacy, and on the other hand, it does not matter if the hypernudges are acceptable to them. These results indicate that what people think is most important, and what people are most concerned about is their privacy. They also pay attention to the inconsistency between the message or suggestion by AI and their feelings at that time. The reason people are upset is the unpredictability of AI results.

DISCUSSION AND CONCLUSION

In this study, we examined whether people accept interventions or are somewhat resistant to AI artefacts. The results indicate that people accept AI support, and they neither proactively choose nor hesitate about the resistance reasons of AI artefacts. Instead, they welcome the support from AI artefacts in many situations. These results may cause a stir as they indicate that users have not deeply considered the acceptance and consequences of AI artefacts. The next issues we will focus on are which categories and types of hypernudges people prefer. However, it is necessary to be alert and deliberate about their unconscious effects before coexisting with AI artefacts hereafter.

One clear limitation of this research is that it might be difficult for participants to imagine what will happen after accepting the AI support. Due to the fact that different types of support provided by AI artefacts has just started, several services are not become familiar to people even though they are rapidly spreading. This suggests that we have to increasingly consider their numerous influences. For example, participants experience many kinds of daily routine tasks on AI artefacts (e. g., smart home, health management apps, and auto-driving system) before the examination (e.g., Dang and André, 2018). The real experience of AI support may be more important when people consider accepting them, such as when customised recommendations are effective or bothersome.

In a similar vein, as this experiment applies the questions considered in prior studies on nudge (e.g., Sustein et al., 2018), hypernudges driven-by AI are not always realistic ones that are prevalent now. It might be due to the fact that the acceptance of hypernudges is relatively high as shown in previous studies. To clarify the cause of the acceptance of AI artefacts, a more practical setting of hypernudges is needed.

There is no doubt that the support services driven-by AI will be ubiquitous and unavoidable for almost all individuals, organisations, and institutions now and in the future. In such situations, we have to consider and judge whether these are acceptable or not. Nevertheless, ‘opacity and incomprehension’ is not selected as the resistance reasons in this study. This might suggest that people neither try nor are eager to understand the AI mechanism, so they are receptive of AI artefacts without deeply considering their effects. Furthermore, the results that the reasons ‘indulging yourself’ and ‘confused whether own will’, and ‘demotivation’ were not chosen by participants adumbrate that these undermined influences make peoples’ choice behaviour unfavourable. This means that people tend to ignore which factors demotivate or pamper them because humans instinctively try to challenge and work earnestly. Therefore, these factors were not selected as reasons for resistance, but they suggest that prior studies on such semiconscious factors might have potential impact on behaviour (e.g., Yamazaki, 2019). Even if AI services are said to be beneficial, they are not always efficient, given that users worry about the consequences of AI artefacts. More academic research is needed to clarify which factors cause anxiety to users.

ACKNOWLEDGEMENTS

This work was supported by the Japan Society for Management Information Grant-in-Aid for SIG 'Monitoring and Control of AI Artefacts', and the Seikei University Grant-in-Aid 2020-22, for the research on 'Monitoring and Control of AI Artefacts: Consideration from Economics, Social, and Legal Perspectives'. I also thank the two anonymous reviewers in ETHICOMP 2021.

KEYWORDS: driven-by AI artefacts systems, acceptance of Hypernudges, privacy, concern for AI artefacts.

REFERENCES

Agarwal, L., Shrivastava, N., Jaiswal, S. and Panjwani, S. (2013). Do Not Embarrass: Re-Examining User Concerns for Online Tracking and Advertising, *Symposium on Usable Privacy and Security (SOUPS)*, July, 24–26.

Andras, P., Esterle, L, Guckert, M., Han, T. A., Lewis, P. R., Milanovic, P., Payne, T., Perret, C., Pitt, J., Powers, S. T., Urquhart, N. and Wells, S. (2018). Trusting Intelligent Machine: Deepening trust within socio-technical systems, *IEEE Technology and Society Magazine*, 37(4), 76-83.

Bhattacherjee, A. and Hikmet, N. (2007). Physicians' Resistance Toward Healthcare Information Technology: A Theoretical Model and Empirical Test. *European Journal of Information System*, 16(6), 725-737.

Chen, F. and Sengupta J. (2014). Forced to Be Bad: The Positive Impact of Low-Autonomy Vice Consumption on Consumer Vitality, *Journal of Consumer Research*, 41, 1089-1107.

Dang, C. T. and André, E. (2018). Acceptance of Autonomy and Cloud in the Smart Home and Concerns, In: Dachselt, R. & Weber, G. (Hrsg.), *Mensch und Computer Tagungsband*. Bonn: Gesellschaft für Informatik e.V. 469-473.

Evers, V., Winterboer, A., Pavlin, G., and Groen, F. (2010). The evaluation of empathy, autonomy and touch to inform the design of an environmental monitoring robot. In Ge, S. S., Li, H., Cabibihan, J.-J., Tan, Y.K. (Eds.), Social Robotics In: Proceedings of the Second International Conference on Social Robotics, [CSR 2010, Singapore, November23-24, (pp. 285-294). Berlin, Heidelberg: Springer Berlin, Heidelberg.

Ferrari, F., Paladino, M. P., and Jetten, J. (2016). Blurring human-machine distinctions: Anthropomorphic appearance, In social robots as a threat to human distinctiveness. *International Journal of Social Robotic*, 8(2), 287-302.

Felsen, G., Castelo, N. and Reiner, P. B. (2013). Decisional enhancement and autonomy: public attitudes towards overt and covert nudges, *Judgement and Decision Making*, 8(3)202-213.

Greene, J. D. and Cohen J. D. (2004). *For the law, neuroscience changes nothing and everything, Philosophical Transactions of the Royal Society of London*. Series B: Biological Sciences, l359, 1775-1785.

Hansen, P. G. and Jesperse, A. M. (2013). Nudge and the Manipulation of Choice: A Framework for the Responsible Use of the Nudge Approach to Behaviour Change in Public Policy, *European Journal of Risk Regulation*, 4(1), 3-28.

Jameson, A. and Schwarzkopf, E. (2002). Pros and Cons of Controllability: An Empirical Study, In Paul De Bra (Ed.). *Adaptive Hypermedia and Adaptive Web-Based Systems: Proceedings of AH 2002.* Berlin: Springer.

Joshi, K. (1991). A Model of Users' Perspective on Change: The Case of Information Systems Technology Implementation. *MIS Quarterly*, 15(2), 229-242.

Jung, J. Y. and Mellers, B. (2016). American attitudes toward nudges, *Judgement and Decision Making*, 11(1), 62-74.

Kane, G. C. and Labianca, G. (2011). IS Avoidance in Health Care Groups: A Multilevel Investigation. *Information System Research*, 22(3), 504-522.

Kang, M. (2009). The ambivalent power of the robot. *Antennae,* 1(9), 47-58.

Kim, T. and Hinds, P. (2006). Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interaction, *The 15th IEEE International Symposium on Robot and Human Interactive Communication*.

Kim, H. W. and Kankanhalli, A. (2009). Investigating User Resistance to Information Systems Implementation: A Status Quo Bias Perspective. *MIS Quarterly*, 33(3), 567-582.

Kirchbuchner, F., Grosse-Puppendahl, T., Hastall, M. R., Distler, M., and Kuijper, A. (2015). Ambient Intelligence from Senior Citizens' Perspectives: Understanding Privacy Concerns, Technology Acceptance, and Expectations, *European Conference on Ambient Intelligence*, 48-59.

Lanzing, M. (2018). Strongly Recommended Revisiting Decisional Privacy to Judge Hypernudging in Self-Tracking Technologies, *Philosophy and Technology*, https://doi.org/10.1007/s13347-018-0316-4. (electronic version)

Lapointe, L. and Beaudry, A. (2014). Identifying IT User Mindsets: Acceptance, Resistance and Ambivalence, *47th Hawaii International Conference on System Science*.

Lapointe, L. and Rivard, S. (2005). A Multilevel Model of Resistance to Information Technology Implementation. *MIS Quarterly*, 29(3), 461-491.

Lewandowsky, S., Mundy, M. and Tan, G. P. A. (2000). The Dynamics of Trust: Comparing Humans to Automation, *Journal of Experimental Psychology: Applied*, 6(2), 104-123.

Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence, *Journal of Consumer Research*, 46, 629-650.

Lorenzi, N. M. and Riley, R. T. (2000). Managing Change: An Overview. *Journal of the American Medical Informatics Association*, 2, 116-124.

Maier, S. F. and Seligman, M. E. (1976). Learned helplessness: theory and evidence, *Journal of Experimental Psychology: General*, 105(1), 3-46.

Marakas, G. M., and Hornik, S. (1996). Passive Resistance Misuse: Overt Support and Covert Recalcitrance in IS Implementation. *European Journal of Information Systems*, 5(3), 208-220.

Martinko, M. J., Zmud, R. W., and Henry, J. (1996). An attributional explanation of individual resistance to the introduction of information technologies in the workplace. *Behaviour & Information Technology*, 15(5), 313-330.

Markus, M.L. (1983). Power, Politics, and MIS Implementation. *Communications of the ACM*, 26(6), 430-444.

Mills, S. (2019). *Into Hyperspace: An Analysis of Hypernudges and Personalised Behavioural Science*, Available at SSRN: https://ssrn.com/abstract=3420211

Pasquale, F. (2015). *The Black Box Society; The Secret Algorithms that Control Money and Information.* Cambridge, MA: Harvard University Press.

Patterson, H. (2013). Contextual expectations of privacy in self-generated health information flows, *TPRC 41: The 41st Research Conference on Communication, Information and Internet Policy*, SSRN: http://ssrn. com/abstract=2242144 or https://doi.org/10.2139/ssrn.2242144.

Peer, E, Egelman, S, Harbach, M, Malkin, N, Mathur, A, Frik, A (2019). *Nudge me right: Personalizing online nudges to people's decision-making styles*.

Raz, J. (1986). The Morality of Freedom, Oxford: Clarendon.

Robey, D. (1979). User Attitudes and Management Information System Use, *Academy of Management Journal*, 22(3), 527-538.

Roubroeks, M., Ham, J., and Midden, C. (2011). When artificial social agents try to persuade people: The role of social agency on the occurrence of psychological reactance. *International Journal of Social Robotics,* 3(2), 155-165.

Sætra, H. S. (2019). When Nudge comes to Shove: Liberty and Nudging in the Era of Big Data, *Technology in Society*, 59, 101130.

Selander, L. and Henfridsson, O. (2012). Cynicism as user resistance in IT implementation, *Information systems Journal*, 22(4), 289-312.

Stein, J., Liebold, B., and Ohler, P. (2019). Stay back, clever thing! Linking situational control and human uniqueness concerns to the aversion against autonomous technology, *Computers in Human Behavior*, 95, 73-82.

Sunstein, C. R. (2016). The Ethics of Influence: Government in the Age of Behavioral Science. CUP, New York.

Sunstein, C. R., Reisch, L. A., and Rauber, J. (2018). A worldwide consensus on nudging? Not quite, but almost, *Regulation & Governance*, 12, 3-22.

Thaler, R. H. and Sunstein, C. (2008): *Nudge: Improving Decisions about Health, Wealth, and Happiness*, New Haven & London: Yale University Press.

Vaes, Paladino, Castelli, Leyens, and Giovanazzi (2003). On the behavioral consequences of Infra-humanization: The Implicit role of uniquely human emotions intergroup relations. *Journal of Personality and Social Psychology,* 85(6), 1016-1034.

Van der Sloot, B. (2017). "Privacy as virtue. moving beyond the individual in the age of Big Data", *School of Human Rights Research Series*, 81.

Verberne, F. M., Ham, J., and Midden, C.J. (2012). Trust in smart systems sharing driving goals and giving information to increase *trustworthiness and acceptability of smart systems in cars. Human Factors*, 54(5), 799-810.

Waytz, A., and Norton, M. I. (2014). Botsourcing and outsourcing: Robot, British, Chinese, and German workers are for thinking not feeling jobs, *Emotion*, l4, 434-444.

Wertenbroch, K., Vosgerau, J., and Bruyneel, S. D. (2008). Free will, temptation, and self-control: we must believe in free will, we have no choice (Isaac B. Singer), *Journal of Consumer Psychology*, 18(1), 27-33.

Yamazaki, Y. (2020). An Empirical Study for the Acceptance of Original Nudges and hypernudges, in Oriva, M., Pelegrin-Borondo, J., Murata, K. and Palma, A. M. (eds.), *Societal Challenges in the Smart Society*, ETHICOMP BOOK SERIES.

Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design, *Information Communication and Society*, 20(1), 118-136.

Zlotowski, J., Yogeeswaran, K., and Bartnec, C. (2017). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources, *International Journal of Human- Computer Studies*, 100, 48-54.

INTEGRATION OF PUBLIC ENGAGEMENT MECHANISMS IN AN ONLINE LANGUAGE COUNSELLING PLATFORM

Urška Vranjek Ošlak, Helena Dobrovoljc

ZRC SAZU, Fran Ramovš Institute of the Slovenian Language (Slovenia),
ZRC SAZU, Fran Ramovš Institute of the Slovenian Language (Slovenia)
and University of Nova Gorica, School of Humanities (Slovenia)

urska.vranjek@zrc-sazu.si; helena.dobrovoljc@zrc-sazu.si

INTRODUCTION

This paper presents an upgrade of the existing software infrastructure of an online language counselling platform with public engagement mechanisms. At the time of the upgrade planning, the platform in question was already well established among users and had great potential to serve as a platform for public-oriented research; however, it was primarily intended to help users with standard language communication. This paper presents steps taken towards consensual integration of the public in the research process through the GRACE project. Activities leading towards an upgrade of the platform started in the autumn of 2019 and are currently in progress. The upgrade will be completed by October 2021.

GRACE aims to contribute to the European Commission goal to spread and embed Responsible Research and Innovation (RRI) in European research, i.e. achieve institutional change. The concept of RRI implies that stakeholders in the research process work together in order to meet the needs of society, namely through democratisation of science, responsiveness and responsibility (*GRACE*, n.d.; Owen et al., 2012). The vision associated with participation in the GRACE project is to develop a plan for defining more advanced forms of collaboration between researchers and the public, and to upgrade the existing platform accordingly. More specifically, we envision a language counselling service that relies on citizen science, based on the perception that the public is competent to conduct carefully structured research in areas that are readily accessible to citizens, such as the language they speak (Svendsen, 2018). Reliance on citizen science will pave the way for the formation of new, more participatory, institutional agendas such as the production of modern language manuals.

Language counselling belongs to the broader spectrum of language management activities (Jernudd & Neustupný, 1987; Lengar Verovnik & Kalin Golob, 2020). In Slovenia, language counselling in various forms represents a notable linguistic tradition. Language counselling activities started in the late 19th century. Today, the so-called language corners (language-oriented newspaper sections), together with popular science language manuals and language counselling forums, fill the linguistic gaps created by the inadequacy of current language manuals (Verovnik, 2016).

CURRENT SITUATION

Language Counselling Platform

The central language counselling platform for the Slovenian language (*Jezikovna svetovalnica*, available at https://svetovalnica.zrc-sazu.si/) managed by Fran Ramovš Institute of the Slovenian Language at Research Centre of the Slovenian Academy of Sciences and Arts (ZRC SAZU) is a good example of how the public can be involved in the scientific process. The platform relies on contributions from the public

in the form of language questions and dilemmas, which are answered by linguists and published on the platform. The ZRC SAZU Language Counselling Service is a reputable consultative body for various language-related professions and for the entire Slovenian language community as it is free of charge and openly accessible; anyone can register and ask questions. The platform receives up to 1,000 views per day and publishes about 30 answers per month. It is widely used for addressing ambiguities in standard language and seeking advice on linguistic choices; researchers use it to identify language description gaps (Dobrovoljc et al., 2020).

The Language Counselling Service is in its core a citizen science service; language counselling cannot be done without the input from the lay public. The lay public is always ready to participate in discussions regarding language use. The Language Counselling Service strives to be democratic in its judgements of language use which agrees with aims and goals of planned activities; the users are invited to participate in order to make the answering process even more democratic. The planned upgrade of the existing platform will not only facilitate citizen science activities; researchers will also benefit from additional user input.

As the Language Counselling Service operates online, it is used by Slovenian language users world-wide. The majority are from Slovenia, followed by users in neighbouring countries (possibly members of minorities). Even though Slovenia is a country with many dialects, the majority of language difficulties refer to standard language.[38] The platform requires users to register in order to ask questions (the purpose of this is to limit user activity to language related questions only). The registration process does not require personal data as users can register using their e-mail, Google or Facebook accounts.

Users of the Platform

Platform users are mostly Slovenian language speakers. They are willing to actively engage in constructive linguistic discussion and research. They perceive standard language as a vital part of their common identity. However, their perception of the linguist's role is not entirely uniform: some are rather reluctant to still accept the traditional role of the linguist as the decision maker (Dobrovoljc, 2004). Others seek straightforward and authoritative linguistic advice on specific issues as well as in-depth explanations of linguistic phenomena. Questions from pupils and students who are still in the process of learning the language are also common.

The platform is a reputable and referential consulting body for laymen and professionals alike; however, language-related professions such as proof-readers, translators, teachers, etc. predominate (Dobrovoljc et al., 2020). Slovenian language users are very interested in their language and its well-being. The need to share their views and opinions on language is very real and frequently expressed. The Language Counselling Service offers competent assistance to users who either face various problems or difficulties that cannot be satisfactorily solved by consulting the available hand-books and manuals, or judge the state of affairs described in the available guides to be in contrast with the actual language use. The Slovenian language has two million speakers who speak either one of the existing eight dialects or one of the regionally spoken language varieties. Mastering the standard language based on the central Slovenian dialect is a challenge, especially for speakers from peripheral regions. In Slovenia, a network of proof-readers and language consultants has developed, and the modern dynamics of language (Internet, social networks, etc.) require up-to-date language manuals. A survey conducted in 2017 (Dobrovoljc et al., 2018; Verovnik, 2016) showed that most questions are asked by

[38] In Slovenian, the standard language is an agreed supra-regional idiom used in the written language since the middle of the 19th century.

users in the 30 to 49 age group. The majority of users have a higher education (university degree, 84%) and the predominant motivation for using the Language Counselling Service is professional need (51%) or the inability to find the answer in available language manuals. Most users indicated that the Language Counselling Service is recognized as a valid reference source in their professional environment.

Language Counsellors

The language counsellors involved are in-house linguists employed by the Fran Ramovš Institute of the Slovenian Language at the Research Centre of the Slovenian Academy of Sciences and Arts who provide language advice as an addition to their professional assignments and scientific research. They sometimes experience assignment overload and therefore have limited time and energy to devote to language counselling. That can sometimes be a problem when many questions are waiting to be answered. To some extent they are sceptical about public participation in the research process. Language users are expected to ask questions and this is approved of. On the other hand, some linguists do not think it is appropriate for the public to express their opinions on linguistic matters.

Also, some believe that language counselling is not particularly valued (it is more the domain of application than research), although some issues demand in-depth research on language dilemmas. Some researchers see language counselling as an unnecessary activity. In this sense, language counselling would need to be re-evaluated as socially and linguistically relevant, particularly in the fields of language policy and science funding.

Existing Public Engagement Mechanisms

Language users are already involved in the upstream stage of the research process as they provide language questions and ambiguities. Language counsellors answer questions of their choice and area of expertise and, after careful evaluation of the Editorial Board, the answers are published on an open platform. At the time of the upgrade planning, the public was not yet officially and in the narrow sense involved in the mid- and downstream stages of the research process; although, occasional feedback via email and the publishing of answers on the platform could be considered downstream public engagement in the broadest sense.

METHODOLOGICAL APPROACH

In the process of identifying the main points where the platform could benefit from a broader aspect of public engagement activities, several steps were taken. First, the Editorial Board of the platform gathered their thoughts and expectations regarding the announced upgrade. Then, a good practice study was conducted by the authors of this contribution to gain insight into similar activities in the field of linguistics. A questionnaire was created to explore the experiences and needs of language counsellors. With all the findings in mind, an upgrade plan was drafted and tested through a consultation process. Through the analysis of the consultation process, the final upgrade plan was devised. In the following subchapters, each of these steps is presented.

Editorial Board Meeting

The Editorial Board meeting revealed that mid- and downstream research stages of the platform have the greatest potential for improvement in terms of public engagement. The platform does not yet

include midstream public engagement activities. There is potential in this direction as the platform reaches a wide audience and has over 1,800 registered users. Editorial Board members were in agreement about including the public in the downstream research stage, namely through the addition of a structured feedback gathering mechanism to the platform.

Study on Good Practices

The study on good citizen science practices in linguistics presented opportunities this methodology enables in the field. Since language is one of the areas of particular interest to the public, language-oriented citizen science activities are likely to be successful in providing large and useful datasets. Most resources (Svendsen, 2018; Stoll, n.d.; *SNF-AGORA*, 2020; *IamDiÖ*, 2020) describe citizen science activities in the up- and midstream of the research process; downstream citizen science activities are less common, which is to be expected given that decision-making in science is usually the preserve of scientists. This was also the case with the involved language counsellors - they too were hesitant when it came to user feedback.

Especially relevant to the platform upgrade activities is the project DiÖ – German in Austria (*IamDiÖ*, 2020), namely its satellite project IamDiÖ which is platform-based in a way that is similar to the ZRC SAZU Language Counselling Service. This project is financed by the Austrian Science Fund. It constitutes of research into the variation and change of the German language in Austria. It explores the use and the subjective perception of the German language in Austria as well as its contact with other languages. The project is institutionally situated at four academic institutions in Austria: University of Vienna, University of Salzburg, University of Graz and the Austrian Academy of Sciences. Citizens of Austria (users of local German varieties) are encouraged to ask questions about their language and to either find answers themselves, or in dialogue with the researchers involved. The lay public submit language related questions (and potentially answer them), gather pictures of writing in public spaces (and potentially analyse them) and create memes.

Here, we focus on the part of the project where the public ask questions about the German language in Austria. These citizen science activities are mainly up- and midstream. The way the public asks questions on language is similar to how the ZRC SAZU Language Counselling Service works (upstream). The submitted questions present valuable research cues. Citizens can conduct their own research as well and tackle language dilemmas. The progress of research and its outcomes are presented in the project blog. The participants are both the lay public in general and language professionals (translators, teachers etc.). The submitted questions are answered by researchers involved in the project, most of them are linguists.

Language Counsellors' Needs and Experiences

A questionnaire was prepared to explore their experiences and needs, and to identify potential mitigation strategies and possible incentives. The questionnaire below consisted of 15 questions addressing these potential issues: work overload, lack of time and energy, unwillingness to accept language users as a vital part of the research process, lack of incentives, lack of appropriate scientific evaluation, and perception of language counselling as an unnecessary activity.

Q1 – In 2019, did you participate in language counselling activities of the ZRC SAZU Language Counselling service?

◯ Yes, I answered at least one language question. (This option allowed the respondent to see all following questions.)
◯ No. (This option would take the respondent to the end of the survey.)

Q2 – Approximately how many language questions did you answer in 2019? (Choose one option.)

◯ less than 5
◯ from 5 to 20
◯ more than 20

Q3 – In your opinion, what are the characteristics of a good answer to a certain language question (e.g. does it describe the situation in language manuals and the language use, does it provide a direct answer to the question, does it provide an evaluation of language alternatives regarding normative adequacy)? (Text.)

Q4 – In your opinion, what is the attitude of Slovenian language users towards language counselling in general? (Text.)

Q5 – In your opinion, are language counselling activities in the ZRC SAZU Language Counselling Service well accepted among Slovenian language users? (Choose one option.)

◯ Yes.
◯ No.
◯ I do not know.

Q6 – In your opinion, is language counselling appropriately valued in science? Please elaborate. (Text.)

Q7 – In your opinion, does participation in language counselling activities put a strain on you? (Choose one option.)

◯ Yes. (This option lead the respondent to question 8.)
◯ No.
◯ Sometimes. (This option lead the respondent to question 8.)
◯ I do not know.

Q8 – Please elaborate. (Multiple choice.)

☐ Answering a language question reveals dilemmas that I did not expect.
☐ It is harder for me to answer questions I do not choose myself, but are instead assigned to me by the Moderator.
☐ I have too little time to participate.
☐ Other:

Q9 – In your opinion, is your participation in the ZRC SAZU Language Counselling Service affected by your primary work tasks? (Choose one option.)

◯ Yes. (This option lead the respondent to question 10.)
◯ No.
◯ Sometimes. (This option lead the respondent to question 10.)
◯ I do not know.

Q10 – Please elaborate. (Text.)

Q11 – What would motivate you to participate in language counselling more often? (Text.)

Q12 – In your opinion, could user feedback on the usefulness of language answers in the ZRC SAZU Language Counselling Service help improve the service quality? (Choose one option.)

◯ Yes.
◯ No.
◯ I do not know.

Q13 – Please elaborate. (Text.)

Q14 – Have you ever used the Q&A database of the ZRC SAZU Language Counselling Service in your research and scientific work, e.g. in your scientific or professional articles or in your lexicographic work? (Choose one option.)

◯ Yes.
◯ No.
◯ I do not know.

Q15 – Please elaborate. (Text.)

Of 23 language counsellors who completed the questionnaire, 20 met the entry condition of having answered at least one language question in 2019. 19 of the 20 respondents completed the survey, 1 respondent completed only the first 6 questions of the questionnaire.

Language counsellors are mainly unanimous that the attitude of language users towards language counselling is positive and that language counselling activities in the ZRC SAZU Language Counselling Service are well accepted among the Slovenian language users. They believe it inappropriate that language counselling is not scientifically evaluated and is not at least partly perceived as a research activity. Even though it is in itself an applied linguistics activity (Orešnik, 1995), it often requires strenuous scientific research. Provided answers to language questions are not included in researchers' bibliographies. Some feel language counselling is perceived as a secondary activity that does not bring research points and is only meant to serve as a promotional activity for the organisation.

Some language counsellors feel language counselling activities sometimes put a strain on them, the predominant reasons being the lack of time, the complexity of language questions and the possible conflict/polemic arising from different views on linguistic matters. Additionally, some language counsellors believe primary work tasks affect or sometimes affect their language counselling activities, mainly through their specialisation; they mostly answer questions related to their field of work. Primary work tasks have priority over language counselling activities.

Language counsellors are highly motivated to answer language questions related to their field of work. Some feel a separate block of time should be reserved for these activities and the work done should have more value and be correctly evaluated. Some feel their participation in language counselling activities would benefit greatly from having a smaller primary task workload.

Language counsellors believe that feedback gathering could be useful for improving the quality of answers to language questions in the spirit of democratization. The possible problem could be that simply gathering feedback on what the user thinks about a certain answer to a language question could

be misleading as language users have very different backgrounds and linguistic knowledge, the motivation behind their questions also differs. Language counsellors believe that the platform's forum, in which the communication between language counsellors takes place, also receives feedback; from other language counsellors, that is. Most language counsellors are in favour of the lay public giving feedback on answers to language question, but they are not in favour of the lay public being able to evaluate language answers in any way in terms of their usefulness or whether the answer was as expected.

As described above, the questionnaire revealed that, next to assignment overload, the lack of formal evaluation for language counselling activities is a topical issue. This problem was communicated to superiors and the search for an appropriate evaluation solution is pending. The main problem, however, proved to be the scepticism of language counsellors towards involving the public in the research process.

The challenge at this stage was to educate the language counsellors who were unwilling to accept public participation about the positive impact of such activities on the research process. The scepticism of language counsellors towards involving the public in the research process was significantly reduced by presenting the findings of the above mentioned good practice study. In addition, a webinar was organised to familiarise the language counsellors with a similar and successful Dutch language portal Meldpunt Taal (represented by Marc van Oostendorp, the portal can be found at http://meldpunttaal.nl/). The webinar consisted of an introduction of the Dutch language portal and its functionalities. The presentation was followed by a lively discussion, mainly about the many similarities between the two platforms. Understandably, the Dutch Language Counselling Service(s) attracted the most attention.

Consultation Process

The aim of the consultation process was to obtain information on how public participation in language research, performed with the help of ZRC SAZU Language Counselling Service, can be increased and how the Service can be improved to meet the needs of users and researchers alike. Through this, the main goal of the consultation process was to test the preliminary upgrade plan.

The consultation process consisted of an online stakeholder consultation organised on Zoom and a questionnaire for lay language users. The stakeholder consultation included three professional language users (a translator, a Slovenian language teacher and a proof-reader) and three researchers, namely established linguists from other research organisations active in language counselling activities. The lay users were invited to complete a questionnaire on the main issues discussed during the online consultation. The questionnaire was published on the platform and was active for one week. The 32 respondents with no linguistic background were predominantly professionals or officials with higher education (mostly BA or MA). The age distribution was quite even in the 36-65 age range. The respondents were not regular users of the Service; they usually use it from a few times a month to a few times a year (or even less often).

The topics discussed were broken down to anticipated upgrade elements of the individual research stages. In the upstream research stage, the possibility of sending language questions by email without registration was considered. Participants agreed that mandatory registration was likely to discourage some platform visitors from asking their questions, but felt that the number of such cases was likely to be small. Language users who seriously want to ask language questions will do so even though they have to register on the platform. Also, allowing unregistered questions would probably lead to an increase in unrelated, irrelevant and incomprehensible contributions.

In the midstream research stage, the possibility of adding an editable text box on the Service's homepage was discussed where news and announcements could be published and which could contain links to midstream research in the future. The participants were enthusiastic about the inclusion of midstream research activities on the platform. They would be willing to participate; incentives could be an additional bonus to attract more users. They felt that the platform had the potential to become a kind of linguistic research community with a limited number of enthusiastic lay linguists.

In the downstream research stage, participants welcomed the prospect of collecting feedback as they felt that this could really improve the quality of the service. They also felt that some ambiguity should be expected in the comments section of the feedback module, as some language users will not be able to explain their opinions coherently or at all. Collecting feedback could also provide information about the quality of the answers and their structure, the comprehensibility of the explanation and even an assessment of how democratic the answers should be.

RESULTS

The results of all the activities listed above were used to create the **final upgrade plan** that describes the public engagement activities and mechanisms that will be integrated into the online language counselling platform in Spring 2021.

Upstream

No changes. Mandatory registration remains.

Midstream

An additional editable section will be secured on the platform homepage where news and announcements will be posted by the platform's Moderator and which could in the future include links to midstream research (language use questionnaires, etc.).

Downstream

There will be two feedback collecting mechanisms – (1) for registered users who ask language questions and (2) for visitors in general. The modules will be separate, as most unregistered visitors "stumble" upon answers on the platform after searching for language advice or solutions to their language dilemmas online, few visit the platform with a specific question in mind.

The collected feedback will be stored in separate databases. Feedback will be collected for every published answer individually. The feedback collecting module for registered users (1) will only be visible to those who ask language questions; they will only provide feedback on answers to their own language questions. The feedback collecting module for platform visitors in general (2) will be visible to all platform guests; visitors in general will be able to provide feedback to any answer they read.

Feedback results and relevant findings will be communicated to language users in several ways: (1) in moderator's replies to topics where individual answers are published, (2) in a circular letter directed towards involved language counsellors, (3) in a notice on the homepage of the platform.

CONCLUSION

Public participation in all research stages of the Language Counselling Service will further democratise the answering process; answering strategies will be adapted to the needs of language users. The platform and the organisation behind it will become more responsive to society, namely by aligning the research process and its outcomes with society's values, needs and expectations as expressed through the public engagement mechanisms described above.

The planned upgrade aims to fully involve the public in the research process. Language users will not only provide research material (upstream), but they will also be able to provide feedback on the research conducted and actively participate in linguistic activities (downstream). In the future, midstream research initiatives could also be published on the platform.

ACKNOWLEDGEMENTS

This contribution is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824521.

We would like to express our gratitude to our project's expert partners: Simeon Veloudis and Maria Michali of SEERC (South-East European Research Centre), and Cristina Paca of ECSITE (The European Network of Science Centres and Museums). Thank you for your support and generous sharing of your knowledge and experiences.

KEYWORDS: linguistics, citizen science, public engagement, language counselling, Internet.

REFERENCES

Dobrovoljc, H. (2004). *Pravopisje na Slovenskem*. Ljubljana: Založba ZRC, ZRC SAZU.

Dobrovoljc, H., Lengar Verovnik, T., & Bizjak Končar, A. (2018). Spletna jezikovna svetovalnica za slovenski jezik ter normativne zadrege uporabnikov in jezikoslovcev. In Bizjak Končar, A., & Dobrovoljc, H. (Eds.), *Zbornik prispevkov s simpozija 2017* (pp. 52-69). Nova Gorica: Založba Univerze. Retrieved from: http://www.ung.si/media/storage/cms/attachments/2018/11/29/16/56/02/Skrabcevi_dnevi_10_11k18.pdf

Dobrovoljc, H., Lengar Verovnik, T., Vranjek Ošlak, U., Michelizza, M., Weiss, P., & Gliha Komac, N. (2020). *Kje pa vas jezik žuli?* Ljubljana: Založba ZRC, ZRC SAZU.

GRACE. (n.d.). Retrieved from http://grace-rri.eu/

IamDiÖ. (2020, April 29). Retrieved from https://iam.dioe.at/

Jernudd, B. H., & Neustupný, J.V. (1987). Language planning: for whom? In Laforge, L. (Ed.), *Actes du Colloque international sur l'aménagement linguistique = Proceedings of the International Colloquium on Language Planning* (pp. 69-84). Québec: Les Presses de L'Université Laval.

Jezikovna svetovalnica. (2021, March 8). Retrieved from https://svetovalnica.zrc-sazu.si/

Lengar Verovnik, T., & Kalin Golob, M. (2019). Spol med družbo, jezikovno rabo in predpisom. *Slavistična revija* 67(2), str. 385-394. Retrieved from: https://srl.si/ojs/srl/article/view/2019-2-1-26

Meldpunt Taal. (2010, September 24). Retrieved from http://meldpunttaal.nl/

Orešnik, J. *Uradi za jezik v Skandinaviji*. Ljubljana: Slovenska akademija znanosti in umetnosti, 1995.

Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy* 39(6), 751-760. Retrieved from https://www.researchgate.net/publication/263662329_Responsible_Research_and_Innovation_From_Science_in_Society_to_Science_for_Society_with_Society

SNF-AGORA Citizen Linguistics (2020, February 12). Zurich Center for Linguistics. Retrieved from https://www.linguistik.uzh.ch/en/forschung/agora.html

Stoll, S. (n.d.). *Citizen Science in Linguistics: Past, Present and Future*. Presentation. Retrieved from https://eua.eu/component/attachments/attachments.html?task=attachment&id=1007

Svendsen, B.A. (2018). The dynamics of citizen sociolinguistics. *Journal of Sociolinguistics* 22(2), 137-160. https://doi.org/10.1111/josl.12276

Verovnik, T. (2016). Jezikovni kotički za sodobno rabo – in sodobnega uporabnika. *Slovenščina danes* 52(7/8), 177-192.

BLOCKCHAIN AND BIOMETRICS AUTHORIZATION; WHAT WE ACTUALLY COUNT TRULY COUNTS?

Kazuyuki Shimizu

Meiji University (Japan)

shimizuk@meiji.ac.jp

ABSTRACT

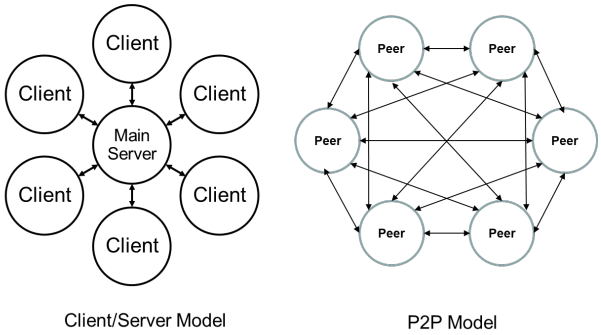
The purpose of this paper is to explore the possibility of sharing value among complexly related stakeholders using blockchain, the core technology of cryptocurrency (Bitcoin) (Nakamoto, 2008). In particular, the value interaction between human input (smell, taste, touch, sound, sight) and human output (tactile, symbol, writing, sound, voice) as IoB (Internet of Body) via the Blockchain is concerned. This interaction is related to the cognitive circuits in the brain (Turnbull, 2012). The dilemma is the autonomous decentralization of systems through P2P (peer-to-peer authorization) and the digital invasion of biological privacy. IoB privacy data is a human rights issue and must be protected appropriately and evaluated. Personal physical data give possibilities for understanding effective predictions of the collective human action. This is like such adopting process of bitcoin into society. In other words, it is similar to finding a causal correlation with a logical way of thinking as the background of data. Does society (collectively cognition) admit new things like Bitcoin as a payment method? Here, the correlation between E-commerce data, Bitcoin prices, Nasdaq and other variables are concerned. As a result, Bitcoin strongly correlated with NASDAQ (.844 **), the regression coefficient with e-commerce data was negative 5%, significantly negative to the Bitcoin price, and 0% strongly positive with the Nasdaq index. (equation; $Y_1=-6.59\times10^{-16}-0.211X_1^{*}+0.068X_2-0.233X_3+0.863X_4^{***}$). It assumed Bitcoin does not work as a major payment method for E-Commerce when the price of Bitcoin continues to rises and fluctuate. However, if Tesla and Visa purchase Bitcoin as their payment method, the price will rise as trustworthy. Therefore, this phenomenon shows that the blockchain technology as a proxy variable the Nasdaq has been very promising.

BLOCKCHAIN

There is legal currency (money) as a method of measuring value. Before the Nixon shock, dollars were proportional to the value of gold (gold standard). As a primary usage, money mediates the exchange of goods and services. However, in the recent Covid19 situation, the value gap between the real economy and the monetary economy is remarkable due to a considerable amount of money supplies from central banks (Piketty, 2014). Gold stands at roughly $10 trillion in 2020 Dec. For example, how long it took a company to be trillion dollars monetary network; Microsoft 44 years, Amazon 24 years, Apple and Google 22 years, Bitcoin 12 years (Michael, 2021). ICOs (Initial Coin Offerings) have emerged in recent years as a new model of funding business venture, in alternative to the traditional investor-mediated capital markets (e.g., Initial Public Offerings (IPOs) and venture capitals).

There are two essential principles in Blockchain. One is the P2P authentication system, and the other is the computational processing speed that depends on the development of technology such as CPU and GPU (graphics processing unit) development. To understand the P2P authentication system, Figure 1 below shows the P2P method and the client-server model.

Figure 1. Client/Server Model versus P2P Model.



Source: Created by the author with reference to Jaime Galán-Jiménez and Alfonso Gazo-Cervero.

Currently, money is a settlement function for mediating various values. Therefore, it can be seen that the function of mediate means using Blockchain, such as Bitcoin, can be used for multiple value exchanges. For example, on YouTube, advertising revenue is determined by the number of followers.

The P2P model is an autonomous decentralized system, and the client-server model is a centralized system. The P2P model shows DAO (Decentralized Autonomous Organization), which realizes democracy in virtual space, see above figure 1.

From this, it can be understood from the comparison that the P2P model is a stable system in which information is not concentrated. A definition of P2P networking is a set of technologies that enable the direct exchange of services or data between computers. Implicit in this definition is the fundamental principles that peers are equals. P2P systems emphasize sharing among these equals. A pure peer-to-peer system runs without any centralized control or hierarchical organization. Cennamo suggests the success of digital currencies depending on their business type (i.e. platform) and on their technology type (i.e., Coin and token) (Cennamo, Marchesi, & Meyer, 2020).

The consensus mechanism that Bitcoin uses is called Proof of Work (PoW). PoW is necessary for security, which prevents fraud, which enables trust. This security ensures that independent data processors (miners) can't lie about a transaction. Blocks are produced roughly every ten minutes and are made up of transactions. Therefore, the miners (block creators) needs in powerful computer equipment like a GPU or, more realistically, an application-specific integrated circuit (ASIC).

To get a sense how much computing power is involved, when Bitcoin launched in 2009 the initial difficulty level was one. As of 2019, it is more than 13 trillion (Euny, 2020).

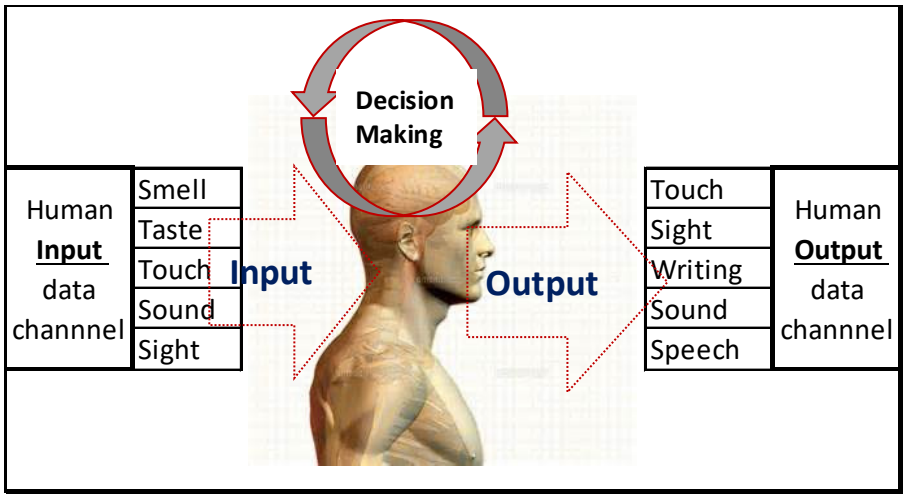
BIOMETRICS

What is the boundary between the Internet of Things (IoT) and IoB? In everyday life, consumers spend about five hours a day on mobile devices and apps such as Facebook, Twitter, and YouTube. The concept of "extended self" has been pointed out by many philosophers. What is a digital identity? We are digitally converted as data via the device. The most recognized example of an Internet of Bodies (IoB) is a pacemaker, a small device placed in the chest to help patients with heart conditions control abnormal heart rhythms with electrical impulses. Increasingly, the external environment of human beings is controlled by smart devices. At the same time, the control from the inside physical body of human beings is added, such as IoB, and the decision making is greatly affected. The promotion of a bioeconomy that contributes to the Paris Agreement and the SDGs sustainable economic growth and

the resolution of social issues is positioned as a national strategy in major countries, and ESG investment by institutional investors is also expanding. (European Commission, 2012)

The human brain is a massively parallel computer processing many bits of data at once rather than one bit at a time. The personal computer could only "emulate about a million neuron connection calculations per second, which is more than a billion times slower than the human brain" (Kurzweil, 2016). Diagram 2 shows the influential factors of human decision-making from input and output as physical actions. When such individual physical activities can be captured as digital data, the interaction between individual autonomy and an organization also could be control.

Figure 2. Concept of Human Transaction.



Source: Created by the author (Shimizu, 2016).

Physiological measurements are made possible, for example, through wearable devices supported by widely used and inexpensive sensors. This technology is currently being introduced to markets. Stream mileage, calories burned, exercise intensity, heart rate and other data to the wearer's mobile device. The Internet of Things (IoT) is built on strong internet infrastructure. Radio Frequency Identification embedded (RFID-embedded) smart cards are ubiquitous, used for many things, including student I.D. cards, transportation cards, bank cards, prepaid cards, and citizenship cards. The development of biometric identification technologies has continued to advance. For example; Blockchain methodologies have been successfully adopted in the library, where based on integrated technologies using Blockchain and finger-vein Biometrics (Meng-Hsuan, 2020 Sep).

Blockchain can offer an autonomous decentralized privacy protection system explained above. Einstein quote, however "Not everything that can be counted counts, and not everything that counts can be counted."

DISCUSSION

This paper tried to examine the possibility the value between stakeholders can be shared using Blockchain. Significantly, the value interaction between human input and output as IoB via the Blockchain was investigated. The dilemma is the autonomous decentralization of systems through P2P and the digital invasion of biological privacy. Murata suggests that in the society where widespread use of artificial intelligence (A.I.)-based information systems, the truths would become meaningless or

worthless. To prevent the emergence of the post-truth society, everyone has to acquire the sufficient knowledge and skill for good computing practices, particularly the ability to consider socially and ethically, through undergoing well-organized ICT educational programmes (Yamazaki, Murata, Orito, Shimizu, 2020). Only human beings can give meaning to "nothing". Does the layered structure provided by platformers in the digital society represent the "meaning" and "value" that humans give? No matter how much you think for, machines can give nothing for this structure. IoB is trying to visualize the part that functions autonomously by our biological data, especially by the autonomic nerves. However, as Murata finally points out, since we give meaning to and understand the numerical values there, we have to give essential meanings that can be seen there by ethical human beings.

IS BLOCKCHAIN TRUSTWORTHY?

There are a lot of works to do for establishing cryptocurrencies, for example, dealing with "Double Spending". Bitcoin can be spent twice as two separate transactions. And "the darknet" side, including illegal porn, settlement for barely legal online gambling, hacking services, and even murder for hire. The anonymous nature of the darknet side for using this newly developed cryptocurrency. (Stephen, 2015) Many problems have to be solved here.

IoB privacy data is a human rights issue and must be protected appropriately and evaluated. Of course, the digital data obtained by IoB can be difficult to identify individuals and track the behaviour of all humans. Still, by clarifying this criterion, it is best for critical personal physical data. It is becoming possible to understand moods and emotions through the regression rates of several variables that contribute to specific behaviours from the movements of the human body using A.I. This is considered to effective predicts the future of collective human action. Today, personal data is valuelessly collected by many for-profit companies such as GAFAM. However, crypto-assets such as Bitcoin can provide a positive or negative solution for this particular situation. Therefore, many for-profit companies, Visa, PayPal, and Tesla, create network effects from new cryptocurrency business models such as PayPal's peer-to-peer wallet CashApp. Apple is also using smartphones for Apple Card fintech with major securities firm Goldman. From similar movements, physical privacy data is traded by crypto assets, for example, to provide valuable physical reaction data to pharmaceutical companies.

Various problems of Bitcoin using such blockchain technology will be solved through the trust of society. Bitcoin adapts the mainstreamer adoption of cryptocurrencies by PayPal, Square, VISA and Tesla. Here we consider the relationship between Bitcoin and the technology companies that should give Bitcoin trust. In other words, the correlation between Nasdaq, an indicator of mainstream tech companies, and Bitcoin prices. As a result, Bitcoin strongly correlated with NASDAQ (.844 **), the regression coefficient with e-commerce data was negative 5%, significantly negative to the Bitcoin price, and 0% strongly positive with the Nasdaq index. The numerical data-target period was from October 2014 to April 2021. Here analyze the correlation and regression rate between monthly E-Commerce transaction data (Y; E-commerce Z) and four variables (X_1), such as Bitcoin Price (X_1: Price Z), the interest rate of Dollar (X_2; U.S. Z)), gold price (X_3; Gold Z) and NASDAQ price (X_4; Nasdaq Z). The monthly E-Commerce transaction data (Y; E-commerce Z) are using the statistical figures of Euromonitor International (BIDITEX Exchange, 2020) (Euromonitor International, 2021). The variable numbers calculated using the Z score to improve the correlation and regression rate analysis accuracy. Therefore, "Z" has written in the table for all variables.

Table 1. Correlation table of four variables.

	Price.Z	US.Z	Gold.Z	Nasdaq.Z	e-commerce.z
Price.Z	1				
US.Z	-0.370	1			
Gold.Z	0.654	-0.791	1		
Nasdaq.Z	0.844	-0.499	0.884	1	
e-commerce.Z	0.717	-0.508	0.877	0.942	1

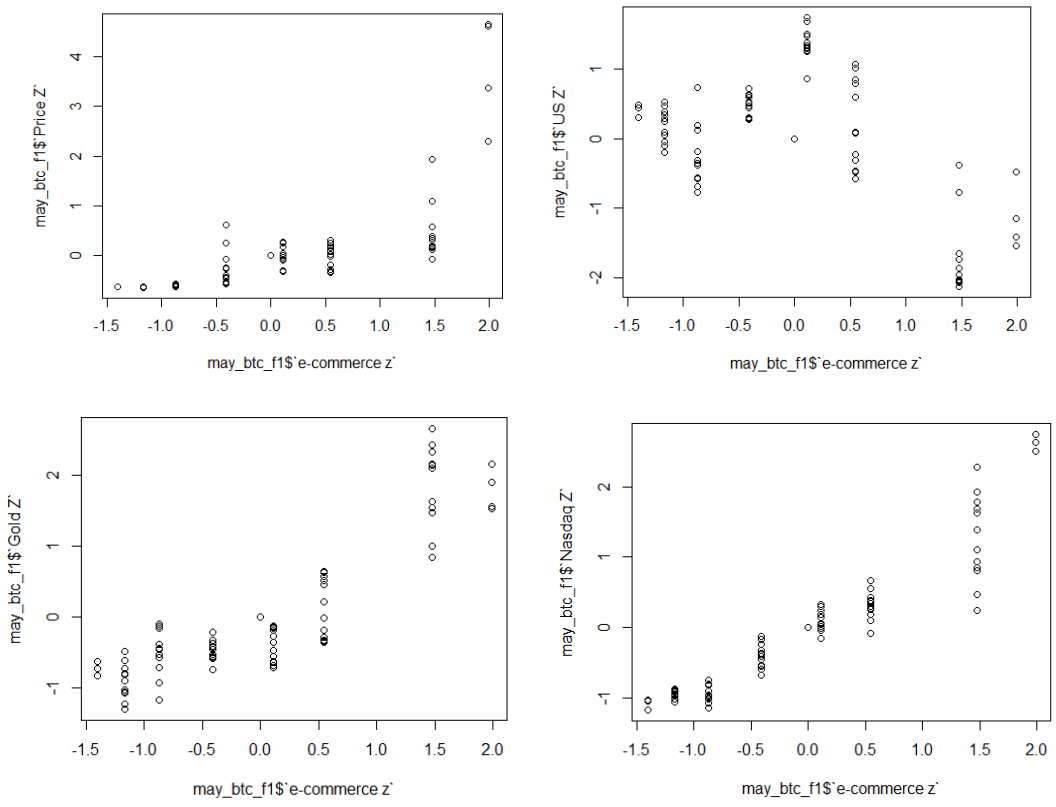
Table 2. Regression table of four variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.59E-15	0.034	-1.92E-13	1.000
`Price Z`	-0.211	0.076	-2.761	0.007*
`US Z`	0.068	0.087	0.781	0.438
`Gold Z`	0.223	0.172	1.300	0.198
`Nasdaq Z`	0.956	0.163	5.862	1.147E-07***

Signif. codes: 0.001 '***' 0.01 '**' 0.05 '*'

equation; $Y_1 = -6.59 \times 10^{-16} - 0.211X_1^{*} + 0.068X_2 - 0.233X_3 + 0.863X_4^{***}$

Figure 3. Scatter plot of four variables.



Source: Created by R.

As a result of Table 1 Correlation rate of four variables, e-commerce data is strongly correlated with Nasdaq (X_4; 0.942), and the U.S. interest rate (X_2) correlates negatively. As a result of Table 2, the regression coefficient with the Bitcoin price is negative 5% significantly, and 0% strongly positive with the Nasdaq index. See figure 1, from the e-commerce and Bitcoin price-scatter plots, it shows an unusual value, so it shows a 5% significant, negative value. It could assume Bitcoin does not work as an alternative payment method for E-Commerce when the price of Bitcoin continues to rises and fluctuate. However, many major fintech companies, Visa, PayPal, and Amazon, seems like creating network effects from these cryptocurrency business models, the price will rise as trustworthy. Therefore, this phenomenon shows that the blockchain technology as a proxy variable the Nasdaq has been very promising. Also, tech-companies were usually inversely correlated with euro and dollar interest rates. This is because government bond yields are preferred over technical investment for certainty during rising interest rates.

KEYWORDS: Blockchain, Peer to Peer (P2P), Biometrics authorization, Decision Making process, Internet of Body (IoB).

REFERENCES

BIDITEX Exchange. (2020, Jan 22). *7 Countries with the Most Bitcoin Hodlers.* Retrieved from https://medium.com/@biditex/7-countries-with-the-most-bitcoin-hodlers-503b205d926f

Cennamo, C., Marchesi, C., & Meyer, T. (2020). *Two Sides of the Same Coin? Decentralized versus Proprietary Blockchains and the Performance of Digital Currencies.* Academy of Management DiscoveriesVol. 6, No. 3Articles. https://doi.org/10.5465/amd.2019.0044

Euny, H. (2020, Nov 18). *Investopedia.* Retrieved from How Does Bitcoin Mining Work?: https://www.investopedia.com/tech/how-does-bitcoin-mining-work/

Euromonitor International. (2021). *Passport*. Retrieved from https://www.portal.euromonitor.com/portal/magazine/homemain

European Commission. (2012). *Innovating for Sustainable Growth: A Bioeconomy for Europe.* Retrieved from https://ec.europa.eu/research/bioeconomy/

Jaime , G.-J., & Alfonso , G.-C. (2011). *Overview and Challenges of Overlay Networks: A Survey.* International Journal of Computer Science & Engineering Survey. Retrieved from https://www.researchgate.net/publication/50199321_Overview_and_Challenges_of_Overlay_Networks_A_Survey

Kurzweil, R. (2016). *The Singularity Is Near: When Humans Transcend Biology (Japanese Edition).* NHK Publishing.

Meng-Hsuan , F. (2020 Sep). Integrated Technologies of Blockchain and Biometrics Based on Wireless Sensor Network for Library. *INFORMATION TECHNOLOGY AND LIBRARIES*. https://doi.org/10.6017/ital.v39i3.11883

Michael, S. (2021). Can Bitcoin Be Replaced or Fail? Retrieved from https://www.youtube.com/watch?v=LdwSxwl9BQk

Moses, A.A., Na, C., Jeng-Shyang, P., Hong-Mei, Y., & Bin, Y. (2020). Securing Fingerprint Template Using Blockchain and Distributed Storage System. *MDPI*. http://doi.org/10.3390/sym12060951

Nakamoto, S. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System.* Bitcoin.org. Retrieved from https://bitcoin.org/ja/bitcoin-paper

Wiener, N. (1950). *The Human Use of Human Beings; Sybernetics and Society.* Retrieved from http://21stcenturywiener.org/wp-content/uploads/2013/11/The-Human-Use-of-Human-Beings-by-N.-Wiener.pdf/ 2014.05.10

Piketty, T. (2014). *Capital in the twenty-first century.* Belknap Press World. Retrieved from https://dowbor.org/blog/wp-content/uploads/2014/06/14Thomas-Piketty.pdf

Shalini, N. (2020, 12 19). *Market Insider.* Retrieved from Bitcoin's market cap could hit $1 trillion in 2021 as its growing reserve currency status drives adoption higher, a cryptocurrency expert says: https://markets.businessinsider.com/currencies/news/bitcoin-market-cap-could-hit-1-trillion-2021-crypto-expert-2020-12-1029908224#:~:text=Bitcoin%20has%20a%20current%20market,stands%20at%20roughly%20%2410%20trillion.

Shimizu, K. (2016). Socio-cybernetic approach into the triumvirate: stakeholder governance between management, shareholders and employees. *An Enterprise Odyssey. International Conference Proceedings*, 273-280. Retrieved from https://search.proquest.com/docview/1815362010?pq-origsite=gscholar&fromopenview=true

Stephen, S. (2015). BITCOIN: THE NAPSTER OF CURRENCY. *J.D., University of Houston Law Center*, 581-641. Retrieved from http://www.hjil.org/articles/hjil-37-2-small.pdf

Turnbull, S. (2012). *A Sustainable Future for Corporate Governance Theory and Practice.* Principal: International Institute for Self-governance. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1987305

Yamazaki, T., Murata, K., Orito, Y., & Shimizu, K. (2020). *Post-Truth Society: The AI-driven Society Where No One Is Responsible.* Universidad de La Rioja: ETHICOMP 2020. Obtenido de https://www.researchgate.net/publication/343099318_Post-Truth_Society_The_AI-driven_Society_Where_No_One_Is_Responsible

PRIVACY IN THE NEW NORMAL: THE IMPLICATIONS OF COVID-19 TRACKING AND TRACING TECHNOLOGIES ON PRIVACY AND CYBERSECURITY

Debora Christine, Mamello Thinyane, Christy Un

United Nations University Institute in Macau (Macau SAR)

debora@unu.edu; mamello@unu.edu; unchristy@unu.edu

ABSTRACT

The COVID-19 pandemic has led governments worldwide to resort to data-driven digital technologies to monitor and curb the spread of the virus. Among these are digital symptom-tracking and contact-tracing technologies. These technologies employ different design approaches with varying potential for intrusive data collection and personal privacy violations. Their deployment has necessitated considering and balancing competing values and addressing ethical concerns in the interest of personal and public safety.

This article positions the deployment of tracking and tracing technologies as a continuation of the pattern of securitized response to the COVID-19 pandemic. It unpacks the privacy concerns in the collection, processing, and use of COVID-19 data vis-à-vis the different forms of privacy breaches. Further, it proceeds to highlight the compatibility and tensions between values central to cybersecurity practices to provide a nuanced understanding of the implications of tracking and tracing technologies on privacy and cybersecurity. In doing so, it seeks to contribute to evaluating the narratives of privacy and cybersecurity in relations to pandemic securitization and providing more nuances to the privacy-security debate.

INTRODUCTION

The COVID-19 pandemic has led governments worldwide to resort to data-driven digital technology-led solutions to monitor and curb the spread of the virus. Despite the established use of surveillance systems to gather information about the spread of infectious diseases, the intensive deployment of surveillance technologies during the COVID-19 pandemic occurs through the mobilization of global health security discourse and securitizing processes in the background. While the deployment of these technologies is built on preexisting global health discourse which frames global health issues, especially disease outbreaks, as constituting risks to national and international stability, the perpetual use and further development of the technology for purposes beyond monitoring of the current pandemic would be facilitated through the mobilization of emergent logics to preempt future public health emergencies.

The securitization of the ongoing pandemic and responses towards it aligns with the widening of the security agenda to include various societal aspects and non-traditional security issues, including climate change, migration, economic recession, and cyberspace, in our world's risk society. According to the classic securitization theory propagated by the Copenhagen School of Security Studies, securitization is a process of presenting a threat as a security threat (securitizing move) to a referent object by a securitizing actor (who performs the securitizing move), who aims to except the referent object from the control of normal politics (Buzan et al., 1998) and rationalize the use of extraordinary measures to secure the object (Williams, 2003; Buzan et al., 1998).

Securitization theory provides a critical lens to examine COVID-19 surveillance technologies as a component of emergent sociotechnical security assemblages used to safeguard public health against the pandemic and provide health security. Further, it provides a framework to contextualize the phenomenon in the increasing techno-securitization of societies which has made surveillance technologies central to securitization and the disciplinary forms of the maintenance of order.

The COVID-19 crisis has indeed posed a fundamental challenge to global security, which extends well beyond the economic and physical security of states. It also posed a substantial threat to human's ontological security, which is one's "fundamental sense of safety in the world" necessary "to maintain a sense of psychological well-being and avoid existential anxiety" (Giddens, 1991: 37). In the face of a heightened sense of uncertainty and vulnerability accompanying the crisis, securitization of the COVID-19 pandemic seeks to restore a sense of ontological security for people by obscuring the political dimension of the global health risk and its governance, providing legitimacy to the integration of policies on public health and public security aimed at countering the threat, and promoting moral imperatives to act collectively.

Within the context of the pandemic securitization, the COVID-19 pandemic is considered an existential threat to the referent object's survival and wellbeing, requiring emergency measures "beyond rules that otherwise bind" (Buzan et al., 1998: 13) – from border closures, travel restrictions, lockdowns, to deployment of surveillance technologies. In this regard, the securitization of the pandemic allows securitizing actors to govern the pandemic through the "state of exception" (Agamben, 2004), which may override the rule of law and infringe upon individual rights (Buzan et al., 1998; Hanrieder & Kreuder-Sonnen, 2014) while conceptualizing the use of digital surveillance technologies as a mechanism of "care" (Newell et al., 2017).

The language of security operates within the "threat-defense" logic and pushes responses to the pandemic away from civil society toward state apparatuses, enabling state encroachment on social life (Buzan et al., 1998). The construction of the nation-states as the referent security object emphasizes a state-centric approach to security and occurs at the expense of an individual's security. A human-centric approach to security, on the other hand, places the protection of individuals' rights and wellbeing at the center of security policies and practices (Deibert, 2018). Following this approach, digital surveillance technologies which allow the excessive collection and intrusive use of data should be designed, deployed, and governed consistent with ethical values and public expectations.

The deployment of digital surveillance technologies, particularly symptom-tracking and contact-tracing technologies, raises privacy concerns not only because of the nature of data generated and how they are extracted, processed, and used. Concerns also revolve around extra technological possibilities created during the pandemic which may remain after the end of the pandemic.

The privacy concerns relating to the security and safety of digital surveillance technologies and data generated from it intersect with the cybersecurity triad in terms of the confidentiality, integrity, and availability (CIA) of COVID-19-related information, technological devices, and information systems. Although the primary ethical motivation of cybersecurity is the prevention of informational harms, its enforcement can also entail harmony and tension with other ethical values.

This article investigates the implications of the deployment of digital surveillance technologies, particularly symptoms-tracking and contact-tracing technologies, against the backdrop of the COVID-19 pandemic securitization, on privacy and cybersecurity. Rather than a focus on the traditional privacy-security dichotomy, it will outline the agreements and tensions between core cybersecurity values regarding the impact of tracking and tracing technologies usage. In doing so, it seeks to contribute to evaluating the changing narratives of privacy and cybersecurity given the pandemic securitization and providing more nuances to the privacy-security debate within the field of digital surveillance.

DIGITAL SURVEILLANCE IN PUBLIC HEALTH

The integration of digital surveillance into the domain of public health operates at different levels – at the individual, interpersonal, national, and global levels. At the global and national level, digital epidemiology and global health security policy discourses and practices resonate with preemptive security logics and the notion of early detection. At the national level, past disease outbreaks have compelled states to explicitly incorporate the protection of the population from emerging lethal infectious diseases into the national security strategy (Elbe, 2009).

At the interpersonal level, doctors exercise personalized health surveillance over their patients. At the individual level, the rise of m-health technologies in the promotion of healthy behaviors and the personalization of health surveillance is one of the trends signifying the integration of digital health surveillance into everyday life. M-health technologies not only enable self-quantification (Lupton, 2016), but also normalize the "datafication" of everyday life.

These developments in digital health surveillance are grounded in global health security discourse on the one hand, and neoliberal politics which focuses on individual behavior and self-responsibility on the other hand. Central to the developments in both epidemiological surveillance and personal health management is an epistemic shift, from reliance on health expert knowledge to a growing reliance on algorithmic knowledge in making sense of information. This demonstrates the emergence of algorithms as a strategic and pragmatic technology of government (Eckmanns et al., 2019).

In the ongoing pandemic, digital surveillance is employed in various forms, from the modeling of disease spread to enforcing the quarantine. The focus of this article is symptoms-tracking and contact-tracing technologies deployed in the COVID-19 pandemic. Symptoms-tracking technologies are syndromic surveillance tools that collect, analyze, interpret, and disseminate health-related data (Berry, 2018). Using these technologies, users can choose to report their symptoms, get a diagnosis, and obtain a triage decision. Tracking is achieved over either Bluetooth or a global positioning system (GPS). The value of symptom tracking technologies lies in the affordance to triage many people.

Contact-tracing was developed based on the infectious disease control strategy of identifying people who may have encountered an infected individual. Contact-tracing uses data from cellular networks (e.g., cellular towers), camera footage, credit card transaction, or smartphones to trace users' exact location (location tracing) and measure their spatial proximity from each other's (proximity tracing). Alert will be sent to all affected users when one user confirms positive for infection. Some applications immediately inform users when an infected user is nearby, thus helping to prevent possible infection by providing timely exposure notification. Contact-tracing also involves regularly following up with contacts to monitor infection (WHO, 2017).

Symptom tracking and contact tracing technologies have long been part of the response to contagious disease outbreaks (Loi et al., 2019). However, to a large extent, they have been conducted manually. Both digital symptom-tracking and contact-tracing technologies provide more accuracies and efficiency in data collection, thus significantly expand the volume and reach of traditional tracking and tracing.

PRIVACY CONCERNS REGARDING TRACKING AND TRACING TECHNOLOGIES

For a long time, state-corporate surveillance has raised questions regarding the invasion of privacy (Cayford & Pieters, 2018). The techno-securitization of the COVID-19 pandemic, however, presents new issues for considering the implications of digital surveillance on privacy, as well as examining the relevance of existing conceptualizations of privacy. Compliance to use digital surveillance technologies,

which can mean agreeing to surrender some forms of privacy, operates within the framing of non-compliance as compromising the security of humanity. The pandemic securitization, therefore, shapes how we define and experience privacy (or loss of it) and security.

Numerous scholars have noted the difficulties of defining the nature and boundaries of privacy. Depending on the accounts from which it is defined, privacy can be understood as a condition, a process, or a goal (Newell, 1998). Normatively, privacy is understood as a value – both as an individual and social value. Legal accounts focus on the right to privacy, while sociological accounts focus on people's interests in the protection of privacy (Gutwirth et al., 2011). Moore (1984) argues that privacy is a "socially created need" for individuals in a society to regulate their interactions with each other to avoid social frictions.

Privacy as an individual value is defined as "the right to be let alone" (Warren & Brandeis, 1890: 193), with an emphasis on seclusion, withdrawal, and avoidance of interactions with others. Privacy as a social value or collective need, on the other hand, refers to "the desire and need of people to come together, to exchange information, share feelings, make plans, and act in concert to attain their objectives" (Bloustein, 2017: 124).

From the perspective of the common good, that is, privacy as a social value, Regan (1995) argues that privacy can be understood as a common value, public value, and collective value. Privacy is regarded as a common value on the basis that all individuals value some degree of privacy and share common perceptions about its importance. The public value of privacy is related to its essence for democracy. The collective value of privacy is best understood through the proposition of the flows of personal information online as a common pool resource.

Despite the various conceptualizations of privacy, privacy debates generally concern acceptable practices regarding accessing, disclosing, and use of personal and sensitive information about a person (Elwood & Leszczynski, 2011). Privacy protects an individual's control over the "acquisition, disclosure, and use" of their personal information against encroaching social control by others (Kang, 1997). In their typology of privacy, Koops et al. (2017) propose informational privacy as a category of privacy that cuts across other categories, i.e., bodily, spatial, communicational, proprietary, intellectual, decisional, associational, and behavioral.

While privacy in the "datafied" society is generally understood as secrecy and data security, it serves a myriad of higher-order social and developmental functions. In this regard, Nissenbaum (2004) posits a concept of contextual integrity (CI) which serves as an appropriate benchmark of privacy. CI binds the protection of privacy to the specific prevailing social and political values and information norms for the gathering and dissemination of information. It rests on the understanding that norms of appropriateness govern people's expectations of how personal information should flow within a given context. In this regard, consideration of harms that may arise from state-corporate surveillance should extend beyond the immediate impact of privacy loss to include moral and political implications of data flows on power structures, justice, fairness, and equality, to determine whether a certain data flow and use of data technology should be allowed or challenged (Nissenbaum, 2010).

CI sets a normative standard in response to the risk of "function creep" and "control creep" across the COVID-19 tracking and tracing technologies. In the context of contact-tracing, for example, people's expectations of the degree of privacy and consent regarding the collection and use of locational and proximity data will be different considering to whom the data flow (e.g., government officials or technology companies) and for what purpose (e.g., for curbing the spread of COVID-19 or long-term governmental use).

Attempts to theorize privacy has yielded, among others, typologies of privacy (see, for example, Koops et al., 2017) and taxonomies of privacy harms (see, for example, Solove, 2006). Solove's taxonomy highlights the multidimensionality of privacy. It allows for an understanding of how a violation of a particular dimension of privacy can lead to a violation of other dimensions of privacy. The taxonomy thus provides a suitable framework to unpack the multitude of ways privacy can be violated beyond misuse of data being collected through the tracking and tracing technologies and the different forms of potential harms produced.

Table 1. The mapping of potential privacy harms of COVID-19 tracking and tracing technologies to Solove's (2006) taxonomy.

Domain	**Privacy breach**	**Potential privacy harms due to approaches in technology use and data process**
Information collection	Surveillance	– Automated, real-time collection of data – Automated reporting to health databases
	Interrogation	Inducing people to divulge personal information
Information processing	Aggregation	Violation of contextual integrity from combining different pieces of information of data subjects from multiple databases
	Identification	– Symptom-based case identification – Re-identification of individuals following anonymization
	Insecurity	Potential data breaches and leakages linked to the centralization of large amounts of personal data
	Secondary use	– Use of data beyond the data subject's consent by data holders – Data removed from their original context of collection
	Exclusion	Data subjects not having legibility to data handling; not having access to aggregated data; not being able to correct errors in automatically collected data; and not having control over data deletion
Information dissemination	Breach of confidentiality	Mishandling - data leaked, breached, or sold to commercial service providers
	Disclosure	– Being discriminated due to one's link to an active cluster – Use of data to justify stigmatization and marginalization
	Exposure	N/A
	Increased accessibility	Higher risk of disclosure and related privacy harms
	Blackmail	Being subjected to extortion and blackmail because of potential data breaches
	Appropriation	Data used for profiling individuals for commercial and policing purposes
	Distortion	– Intentional manipulation of data – Inaccurate reporting as infected persons or linked to a cluster
Invasion	Intrusion	– Altering users' behavior into complying with COVID-19 guidelines through nudging and persuasive techniques – Alerting authorities and citizen vigilantes about those violating quarantine rules
	Decisional interference	– Altering users' behavior into complying with COVID-19 guidelines through nudging techniques – Enforcing curbing measures through geofencing techniques – Mandatory use without opt-in and opt-out choices

Source: authors' elaboration based on the functionalities of COVID-19 tracking and tracing technologies (Ada Lovelace, 2020; WHO, 2020)

Information collection

Privacy violation at the stage of information collection occurs due to the means through which information about an individual is obtained. The pervasiveness of data collection in the datafied society means that people are subject to intensified scrutiny whose daily lives are captured as data. As submission into the datafication regime becomes a precondition for integration into society, the notion of consent remains central to determining if privacy is breached during the datafication process.

Most contact-tracing technologies automatically and in real-time mode collect the location data of mobile phone users via mobile networks or Bluetooth, without options to opt-in and opt-out (Sowmiya et al., 2021). Like Aarogya Setu in India, some require users to input their name, mobile number, age, gender, profession, and details of countries they have visited in the last 30 days. The rationale for automated contact-tracing is to expand the volume and reach of traditional contact-tracing. It, therefore, provides the ability to automatically trace the intersections of millions of people, including those who have retrospectively been in contact with infected people, and potentially limit additional waves of infections.

Surveillance, in the form of continuous monitoring of symptoms, movement, and social interactions, functions as social control for recording acts of compliance and deviance in the pandemic time. Such continuous monitoring remains problematic for its inhibitory effects, which may lead people to perform self-censorship.

Interrogation is an integral part of contact tracing. It involves interviews of suspected and confirmed COVID-patients and their contacts to ascertain who they have been in contact with, to identify other infected people before they infect more people. Both self-response questionnaires given to people and contact-tracing interviews can be interrogation methods to compel them to divulge information against their will. Interrogations occur within the bound of the rationales of pandemic securitization, and some degree of coerciveness is involved in it.

Information processing

Privacy concerns at this stage are related to the way data is handled after collected. Key issues of data handling generated from contact-tracing technologies include who gets access to data, methods for data analysis and interpretation, data retention period, and the purposes for which data would be used during and after the pandemic.

Data **aggregation** refers to analyzing and presenting data at the aggregate level. This is performed to obtain critical insights into the pandemic for modeling and prediction efforts. For reducing the risk of revealing personal information about individuals' lives, data aggregation is regarded as useful to preserve individual privacy. However, data aggregation involves combining bits and pieces of personal data from multiple databases, disconnecting information from the original context in which it is gathered, thus potentially violating privacy as contextual integrity (Nissenbaum, 2004, 2010). Concerns about the veracity and reliability of the voluminous streams of real-time symptom-tracking and contact-tracing data are also central issues to privacy violation through aggregation. This big data, despite its high coverage, often have multiple gaps, biases, errors, and inconsistencies (Kitchin, 2014), risking inaccurate, unfair representation of individuals.

Aside from telecommunications transactions, financial transactions data, social media analytics (Gitzen, 2020), personal health informatics (Abuhammad et al., 2020), and personal information collected in different settings are combined to generate more accurate population-level mobility data and track the progression of COVID-19 and its symptoms. Further, COVID-19 symptom-tracking sees

an increasing device-driven approach following the deployment of wearables, including smartwatches and activity trackers. While users of wearables are comfortable sharing their data with the wearable developers to receive personalized health and wellbeing-related service, the data flow to the government for pandemic management and beyond is a new information flow of which assessment of appropriateness should be contextually dependent.

According to the privacy as CI framework, such evaluation should be conducted based on the following informational norms (Nissenbaum, 2010): context (the securitization of the pandemic that governs the interactions of government and individual citizens), actor (the government as users of aggregated data and individual citizens as data subjects as well as the objects upon which disciplinary actions are imposed), attributes (the types of personal information in play), and transmission principles (the specific constraints regulating the information flow from the different data settings to government use for pandemic management).

Contact-tracing inherently violates privacy through its affordance to link streams of data to individuals, both COVID-19 infected persons and those who have been in contact with them. In some regions, the identity of positive cases is published on the government's website (Singer & Sang-Hun, 2020). Further, data aggregation and anonymization have proven to be insufficient in assuring users' privacy due to the possibility for re-identification through combing and combining datasets (de Montjoye et al., 2013; Narayanan & Shmatikov, 2009). Privacy breach through **identification** is increasingly straightforward as extensive aggregations of data about an individual across many databases are deployed towards preventing the spread of COVID-19 and predicting high-risk zones. The South Korean case of the link of the queer communities to a COVID-19 outbreak provides a clear example of how identification as a form of informational privacy violation can lead to discrimination. Meanwhile, identification in the form of linking individuals to a cluster of protest outbreaks, either through symptom-tracking or contact-tracing, could function as a tool to silence dissent and round up disfavored citizens.

Insecurity in the context of tracking and tracing is particularly related to data storage. Both symptom-tracking and contact-tracing technologies either use a centralized or decentralized approach to data storage. In centralized systems, data is stored on government servers, while in decentralized systems, most data are stored on users' phones. While the government's servers may be protected by stronger cybersecurity measures, centralized tracking and tracing databases are "honeypots" to adversaries, thus risks harming more people in the case of data leak or breach. Both approaches, however, have the potential for data to be leaked or improperly accessed via the interactions between mobile applications in a user's phone, between phones of users nearby, and via communications with central servers (Angwin, 2020).

Since the beginning of the rollout of COVID-19 surveillance technologies, there have been widespread concerns regarding their affordances to enable a new biopolitical architecture through the expansion of the use of COVID-19 data and data aggregated during the pandemic. These data are subject to control creep, their original purpose is being extended from implementing disciplining and control forms of governmentality in the context of pandemic management (e.g., nudging people to actively comply with social distancing) to perform surveillance, governance, and predictive policing for political ends to monitor and discriminate against certain groups. There are also privacy concerns regarding the repurposing and monetization of tracking and tracing data and data aggregated from it by private enterprises without the consent of the users of the technologies.

Secondary uses of COVID-19 create dignitary harm for individuals as such uses impede people's expectations about how data they consent to share will be used. Secondary uses also violate privacy

as CI as such uses remove data from the original context in which it is collected, thus risking data being misunderstood.

The **exclusion** of data subjects from getting informed about and participating in the handling and use of their data is particularly related to their illegibility to personal data collection and processing, lack or no access to aggregated data, and control over the deletion and retention of their data. Shutting out people's participation in the COVID-19 data processing and use harms their privacy by creating a sense of insecurity and distrust about the data systems and loss of control of their data.

Information dissemination

Privacy harms in the information dissemination stage concern the revelation of personal data or the threat of spreading information. **Breach of confidentiality** in the context of the COVID-19 pandemic occurs through pandemic securitization rationales and the lack of capacity of data holders in providing data security. To support contact tracing, confidential COVID-19 patient information has been made public on the government's website in several regions without the consent of individuals. Further, some COVID-19 contact tracing breach cases have surfaced concerning mismanagement of vendors and direct attacks on data holders.

The disclosure of COVID-19 personal and sensitive personal data affects the way society views an individual and the groups the individual is associated with. **Disclosure** of a link to a COVID-19 outbreak or positive cases resulting from contact tracing, for example, could lead to people, particularly vulnerable groups, being discriminated against. In South Korea, the detailed embarrassing revelations of the places visited by COVID-19 infected individuals' before testing positive have not only inhibited their associational privacy (Koops et al., 2017) but also threatened their security. Such information can also be used by others to highlight the deviance of specific groups from societal norms and justify their stigmatization and marginalization.

Making obscure COVID-19 tracking and tracing data more accessible (**increased accessibility**) enhances the risk of disclosure, and therefore the privacy harms related to disclosure. The publication of detailed information of infected individuals on the government website, for example, produces more risk of disclosure for the individuals and their contacts compared to an automated alert sent to those with who they have been in close contact.

The sensitive nature of COVID-19 data obtained through tracking and tracing applications, as well as the amount of personal data aggregated and stored to diagnose, recognize trends, and to manage the pandemic, risks the use of data by criminals to blackmail individuals and of the victims of blackmail to give in to the threat. **Blackmail** allows criminals as data holders to exercise control and domination over individuals as data subjects.

The privacy concerns of appropriation are related to the impingement of an individual's freedom in the presentation of their identity to serve the aims and interests of another. **Appropriation** is closely related to disclosure in the way individuals lose control over their self-representation in public. Appropriation occurs in the case where personal identifiers, activity data, mobility data, and contacts and affiliations are used to profile individuals, both for commercial and predictive policing purposes. The aggregation of data collected from symptom trackers with activity data and mobility data, for example, could help health service providers and tech companies discover new products and markets. Similarly, several cases revealing the use of contact-tracing data by the government to profile dissents and link them to the increase of COVID-19 infections reflects how data appropriation breaches an individual's privacy.

Distortion is closely related to disclosure as both concern the impact of individuals' lack or loss of control over the dissemination of their personal information on the interpretation of those information pieces. The revelation of one's COVID-19-related information can harm people due to intentional manipulation of that information or inaccuracy in information collection and processing. While the former is related to the malicious use of COVID-19 data, the latter is due to technological or methodology issues. The dissemination of information of false positives, for instance, not only gives a false sense of insecurity but could damage an individual's associational privacy and reputation.

Exposure involves the revelation of an individual's nudity, grief, or bodily functions. In the context of COVID-19 pandemic, this privacy breach does not necessarily result from the use of tracking and tracing technology.

Invasion

Intrusion involves "invasions or incursions into one's life" (Solove, 2006: 549) while **decisional interference** involves interference to an individual's autonomy. Both symptom-tracking and contact-tracing technologies could potentially disturb people's tranquillity and solitude through their affordances of continuous surveillance of people's mobility and activities, as well as constant exposure notification alerts and reminders that they provide users. Further, for the feature which alerts authorities and other users about those violating quarantine rules, both technologies are designed to give users discomfort for their non-compliance to COVID-19 public health measures.

Through their affordance to nudge users' behaviors and enforce COVID-19 public health measures, these technologies can create a chilling effect on individuals' decisions regarding their bodily, spatial, and behavioral privacies (Koops et al., 2017). Further, mandatory use of the technologies without the opt-in and opt-out options essentially interferes with individuals' decisions regarding their healthcare and technology use.

CYBERSECURITY VALUE TENSIONS

While the relations between cybersecurity and privacy have long been noted, their relations have largely been framed as a dichotomy. However, the cybersecurity-privacy trade-offs approach is insufficient to discuss the ethical complexity of privacy and cybersecurity within the current context of pandemic securitization. Not only the privacy versus security framing fails to account for the nuanced interactions between the two values – which can be conflicting and reinforcing depending on the context – it is also too one-dimensional in that it ignores other values that are at stake in cybersecurity. This is because, in some cases, even if privacy and security concerns are addressed, individually or all at once, there could still be ethical issues that need addressing in the deployment of digital technology beyond privacy and security.

This section focuses on the implications of the tracking and tracing technologies on cybersecurity through exploring the conflicts and harmonies between core cybersecurity values - security, privacy, fairness, and accountability (Christen et al., 2020). Security can be understood as the state of being free from threats; fairness is related to equality, justice, and non-discrimination; and accountability concerns transparency, openness, and explainability.

Privacy and security can strengthen or weaken each other depending on the context. Privacy-preserving measures integrated into tracking and tracing technologies, for example, could provide individuals with both privacy and cybersecurity because safeguarding individual's data privacy requires

some degree of cybersecurity. Meanwhile, some degree of privacy in the COVID-19 data achieved through aggregation, anonymization, and encryption contributes to individuals' cybersecurity by minimizing the risk of identification. Most data generated through tracking and tracing technologies are stored in government servers, enabling them greater control over data handling, storage, and use. While these servers usually have better cybersecurity protection compared to users' phones, they appear as "honeypots" to adversaries due to the amount and nature of data stored. Such a centralized approach thus has a greater risk for harming individuals' privacy.

Instead of being inevitable, the contentions between privacy and security occur within the context of the clash of notions of security: between the state-centric and human-centric approaches to security. A state-centric approach to security considers the state as the referent object that needs securitizing while a human-centric approach considers humans as the referent security object. A human-centric approach to security considers individual privacy protection as a component of human security (Deibert, 2018). It emphasizes the importance of "distributed security" instead of "centralized security". The approach recognizes the need to translate citizen's control over their data through a real authorization of control and legal consequences for those violating citizen's data privacy. On the other hand, a state-centric approach to security focuses more on protecting the security of the state in favor of humans and considers increasing the security of individual citizens decreases the state's security. This neglect of the human element translates to individual's sustained feelings of insecurities concerning their vulnerabilities in the infosphere (Dunn Cavelty, 2014).

A human-centric approach to the pandemic securitization would ensure the operationalization of a privacy-first approach to the design and deployment of tracking and tracing technologies as well as COVID-19 data flow in promoting a comprehensive interpretation of human security (see, for example, MIT Media Lab, 2020).

The clash of the notions of security is echoed in the security-fairness value conflicts. **Security** may conflict with **fairness** when its measures risk people's access to fair processes and outcomes of COVID-19 data systems. Evaluation of fairness of the implementation of tracking and tracing technologies and the pandemic data systems therefore should not only be done from the viewpoint of state-centric security but also human-centric security.

The **privacy-fairness** relations can be framed from the perspective of data justice – "fairness in the way people are made visible, represented, and treated" in data and data-related outcomes (Taylor, 2017). Fairness and privacy conflict in at least two instances. First, digital inequality in terms of access to digital technology and digital skills can result in imbalances of representation in COVID-19 data in terms of demographics, geography, and socioeconomic vulnerabilities. Those who are not represented or misrepresented in COVID-19 data systems due to technical limitations of the information and communication architecture, including imprecision of the tracking and tracing technologies, can be excluded from enjoying pandemic relief supply. While lack of representation in COVID-19 data systems might mean more privacy, it could also inhibit at-risk individuals from getting the support they need and are entitled to. Secondly, automated contact-tracing is more likely to return false positives for marginalized populations and vulnerable communities in relatively crowded, high-contact conditions with fewer resources to self-isolate, thus risking compounding historical discrimination.

Privacy supports fairness in the way that minimization of data collection could impede individuals from being treated unfairly due to their personal information. Because the risks of exposure to practices of governmentality are unevenly distributed across the population, less representation in COVID-19 data systems could result in more fairness for marginalized groups.

Accountability measures such as imposing strict purpose, time, and access limitations to data, and establishing an independent oversight to scrutinize technology deployment and data use could

contribute to preserving individuals' privacy. However, some degree of revelation of personal information which might be privacy-sensitive, such as socioeconomic status, might be needed to obtain the bigger picture of the effectiveness of the pandemic handling.

Individuals' exclusion from their participation in the handling of personal data that is collected and used about them through tracking and tracing technologies reduces the accountability of the COVID-19 data systems. This lack of accountability goes results in individuals' sense of insecurity. A state-centric approach to security, on the other hand, would view that too much **accountability** undermines the effectiveness of (cyber)**security** measures embedded in the pandemic data systems.

CONCLUSION

The assemblages of COVID-19 pandemic securitization have enabled the mobilization of rationalities, epistemic, technological and data practices, as well as governmentalities that support the deployment of COVID-19 digital surveillance technologies to achieve a state-centric notion of security in favor of human-centric security, particularly in the form of individual privacy. This article has reviewed the myriad potential privacy breaches resulting from the deployment of symptom-tracking and contact-tracing technologies in the COVID-19 pandemic and the COVID-19 data systems. The identified informational privacy breaches can trigger interrelated privacy issues across dimensions of reputational, associational, behavioral, decisional, and bodily privacy. It is also found that the collection, processing, and dissemination of COVID-19 data contribute to a greater power imbalance between individual citizens with the state and tech companies as data holders and data users – an issue Solove (2006) terms as "architectural" problem. This imbalance enhances the risk of abuse of power.

The article has contextualized the discussion about the harmonies and tensions between core cybersecurity values within the pandemic securitization to consider the impacts of COVID-19 tracking and tracing technologies on cybersecurity. We argue that evaluation of security, and by extension, cybersecurity, should be done both from the perspective of state-centric security and human-centric security. Balancing the two notions of security is a path towards satisfying the need for protecting individual privacy and safeguarding public health and security, both in normal times and during crises.

ACKNOWLEDGEMENTS

This work is supported by the Science and Technology Development Fund of Macau (FDCT) under Grant No. 0016/2019/A.

KEYWORDS: privacy, security, technology, COVID-19, cybersecurity, value.

REFERENCES

Abuhammad, S., Khabour, O. F., & Alzoubi, K. H. (2020). COVID-19 Contact-Tracing Technology: Acceptability and Ethical Issues of Use. *Patient Preference and Adherence*, *14*, 1639-1647. https://doi.org/10.2147/PPA.S276183

Ada Lovelace. (2020). *Exit through the App Store? Rapid evidence review*. https://www.adalovelaceinstitute.org/case-study/exit-through-the-app-store/

Agamben, G. (2004). *State of Exception* (K. Attell, Trans.). University of Chicago Press. https://philpapers.org/rec/ATTSOE

Angwin, J. (2020, April 14). Will Google's and Apple's COVID Tracking Plan Protect Privacy? *The Markup*. https://themarkup.org/ask-the-markup/2020/04/14/will-googles-and-apples-covid-tracking-plan-protect-privacy

Berry, A. C. (2018). Online Symptom Checker Applications: Syndromic Surveillance for International Health. *The Ochsner Journal*, *18*(4), 298-299. https://doi.org/10.31486/toj.18.0068

Bloustein, E. J. (2017). *Group Privacy: The Right To Huddle*. Routledge. https://www.taylorfrancis.com/chapters/mono/10.4324/9781351319966-4/group-privacy-right-huddle-edward-bloustein

Buzan, B., Wæver, O., & de Wilde, J. (1998). *Security: A New Framework for Analysis*. Lynne Rienner Publishers.

Cayford, M., & Pieters, W. (2018). The effectiveness of surveillance technology: What intelligence officials are saying. *The Information Society*, *34*(2), 88-103. https://doi.org/10.1080/01972243.2017.1414721

Christen, M., Gordijn, B., & Loi, M. (Eds.). (2020). *The Ethics of Cybersecurity* (Vol. 21). Springer International Publishing. https://doi.org/10.1007/978-3-030-29053-5

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*(1), 1376. https://doi.org/10.1038/srep01376

Deibert, R. J. (2018). Toward a Human-Centric Approach to Cybersecurity. *Ethics & International Affairs*, *32*(4), 411–424. https://doi.org/10.1017/S0892679418000618

Dunn Cavelty, M. (2014). Breaking the Cyber-Security Dilemma: Aligning Security Needs and Removing Vulnerabilities. *Science and Engineering Ethics*, *20*(3), 701-715. https://doi.org/10.1007/s11948-014-9551-y

Eckmanns, T., Füller, H., & Roberts, S. L. (2019). Digital epidemiology and global health security; an interdisciplinary conversation. *Life Sciences, Society and Policy*, *15*. https://doi.org/10.1186/s40504-019-0091-8

Elbe, S. (2009). Virus Alert. In *Virus Alert*. Columbia University Press. https://www.degruyter.com/document/doi/10.7312/elbe14868/html

Elwood, S., & Leszczynski, A. (2011). Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum*, *42*(1), 6-15. https://doi.org/10.1016/j.geoforum.2010.08.003

Favaretto, M., De Clercq, E., & Elger, B. S. (2019). Big Data and discrimination: Perils, promises and solutions. A systematic review. *Journal of Big Data*, *6*(1), 12. https://doi.org/10.1186/s40537-019-0177-4

Giddens, A. (1991). *Modernity and Self-Identity: Self and Society in the Late Modern Age*. Stanford University Press. http://www.sup.org/books/title/?id=2660

Gitzen, T. (2020, June 18). Tracing homophobia in South Korea's coronavirus surveillance program. *The Conversation*. http://theconversation.com/tracing-homophobia-in-south-koreas-coronavirus-surveillance-program-139428

Gutwirth, S., Poullet, Y., De Hert, P., & Leenes, R. (Eds.). (2011). *Computers, Privacy and Data Protection: An Element of Choice*. Springer Science & Business Media. https://doi.org/10.1007/978-94-007-0641-5

Hanrieder, T., & Kreuder-Sonnen, C. (2014). WHO decides on the exception? Securitization and emergency governance in global health. *Security Dialogue*, *45*(4), 331-348. https://doi.org/10.1177/0967010614535833

Kang, J. (1997). Information Privacy in Cyberspace Transactions. *Stanford Law Review*, *50*(4), 1193-1294.

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE.

Koops, B.-J., Newell, B. C., Timan, T., Chokrevski, T., & Gali, M. (2017). *A Typology of Privacy*. *38*, 93.

Loi, M., Christen, M., Kleine, N., & Weber, K. (2019). Cybersecurity in health – disentangling value tensions. *Journal of Information, Communication and Ethics in Society*, *17*(2), 229-245. https://doi.org/10.1108/JICES-12-2018-0095

MIT Media Lab. (2020, April 10). *Safe Paths: A privacy-first approach to contact tracing*. MIT News | Massachusetts Institute of Technology. https://news.mit.edu/2020/safe-paths-privacy-first-approach-contact-tracing-0410

Moore, B. (1984). *Privacy: Studies in Social and Cultural History*. https://www.routledge.com/Privacy-Studies-in-Social-and-Cultural-History/Moore-Jr/p/book/9781138045262

Narayanan, A., & Shmatikov, V. (2009). De-anonymizing Social Networks. *2009 30th IEEE Symposium on Security and Privacy*, 173-187. https://doi.org/10.1109/SP.2009.22

Newell, B. C., Gomez, R., & Guajardo, V. E. (2017). Sensors, Cameras, and the New 'Normal' in Clandestine Migration: How Undocumented Migrants Experience Surveillance at the U.S.-Mexico Border. *Surveillance & Society*, *15*(1), 21-41. https://doi.org/10.24908/ss.v15i1.5604

Newell, Patricia B. (1998). A cross-cultural comparison of privacy definitions and functions: A systems approach. *Journal of Environmental Psychology*, *18*(4), 357-371. https://doi.org/10.1006/jevp.1998.0103

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review*, *79*(1), 119-157.

Nissenbaum, H. (2010). *Privacy in context: Technology, policy and the integrity of social life*. Stanford Law.

Regan, P. M. (1995). *Legislating Privacy: Technology, Social Values, and Public Policy*. University of North Carolina Press. https://www.jstor.org/stable/10.5149/9780807864050_regan

Solove, D. J. (2006). A Taxonomy of Privacy. *University of Pennsylvania Law Review*, *154*(3), 477. https://doi.org/10.2307/40041279

Sowmiya, B., Abhijith, V. S., Sudersan, S., Sakthi Jaya Sundar, R., Thangavel, M., & Varalakshmi, P. (2021). A Survey on Security and Privacy Issues in Contact Tracing Application of Covid-19. *SN Computer Science*, *2*(3), 136. https://doi.org/10.1007/s42979-021-00520-z

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society*, *4*(2), 205395171773633. https://doi.org/10.1177/2053951717736335

Warren, S. D., & Brandeis, L. D. (1890). The Right to Privacy. *Harvard Law Review*, *4*(5), 193-220. https://doi.org/10.2307/1321160

WHO. (2017). *Infection prevention and control: Contact tracing*. https://www.who.int/news-room/q-a-detail/contact-tracing

WHO. (2020). *Digital tools for COVID-19 contact tracing* (WHO/2019-nCoV/Contact_Tracing/Tools_Annex/2020.1; COVID-19: Surveillance, Case Investigation and Epidemiological Protocols). https://www.who.int/publications-detail-redirect/WHO-2019-nCoV-Contact_Tracing-Tools_Annex-2020.1

Williams, M. C. (2003). Words, Images, Enemies: Securitization and International Politics. *International Studies Quarterly*, *47*(4), 511-531.

EVOLUTION IN THE MUSEUM NETWORK AND ITS USE IN THE COVID-19 PANDEMIC

Raquel García-Martín, Ana María Lara Palma, Bruno Baruque Zanón

Universidad de Burgos (Spain)

rgm0004@alu.ubu.es; amlara@ubu.es; bbaruque@ubu.es

INTRODUCTION

The city of Burgos, in the north of Spain, is home of one of the few museums in the world dedicated to Human Evolution: The Human Evolution Museum (Museo de la Evolución Humana, MEH). It was originally created as an environment to show to the public the achievements of the archaeological research in the Atapuerca archaeological sites (town of the province of Burgos, 15 kilometers from the capital) but it has evolved to a landmark to the city, becoming also central to its international attractive and cultural life.

The numbers of fossils found was increasing, as was their research and their importance. Hence the need to create a museum and research center to investigate, conserve, inventory and expose the material found in the Atapuerca sites. Its objectives include accessibility, and promotion of awareness and knowledge transfer to all types of public, children, youth and adults, from schoolchildren to expert people, national and international or people with disabilities.

In a very schematic way, it can be equated to chain of knowledge, which begins in excavations, is studied in the research center, is exhibited in the museum and transcends the public (general public and/or scientific community), both fossils, as the results of the studies carried out. All this by different means, physical, face-to-face or virtual.

It´s a great potential that is in the province of Burgos with the possibility of visiting the sites and the museum in the same province where then findings that are developed during the annual excavations have been found and where the investigations are carried out in the Research Center National Human Evolution (*Centro Nacional de Investigación sobre la Evolución Humana,* CENIEH) (Alonso Alcalde, 2018).

The international importance is demonstrated by some of its achievements and accolades. Some of them are the one obtained by the Atapuerca Research Team (Equipo de Investigación Atapuerca, EIA) in 1997, which received the Prince of Asturias Award for “Scientific and Technical Research” and in 2000, the Atapuerca archaeological sites received the qualification of “World Heritage Site “by UNESCO, as a consequence of the exceptional archaeological findings, which it houses inside, among which the fossils of four different species of hominids stand out. Becoming one of the most important sites in Europe, due to the amount of remains located in the same area, which makes it a unique place to meet our ancestors and an essential topic for any study on human evolution (Alonso Alcalde, 2018).

The work in common and the same direction, of different centers such as the MEH an CENIEH, with different private and governmental entities, where the Atapuerca Foundation and Atapuerca Culture of Evolution System (Sistema Atapuerca Cultura de la Evolución, SACE) carry out coordination tasks between them to administrative, management, and research level, although with different functions and manager each of them.

The objective of this paper is to observe the positive impact on society, for the acquisition and accessibility of information and data, through “Cultural Digitization” in museums and to the way of using New Technologies. Knowledge Management (KM) gives importance to the generation and transfer of knowledge as well as to technology (Correa Drummond de Alvarenga Neto & Gomes Vieira, 2011).

THEORICAL FRAMEWORK

Although the theme of the Museum revolves around the evolution of man since prehistory, we find ourselves in the 21st century, in a relatively young museum -cutting edge in many aspects- and a reference scientific dissemination center, where digital technology, internet, and social networks are very present.

The digital media factor must now be considered, as they are increasingly used, in aspects such as online training and / or dissemination, and through different channels, such as social networks or digital platforms. Information technology-based knowledge strategy influences day-to-day work performance and scientific production in research centers and universities (Fernández-López et al., 2018).

From the point of view of the application of new technologies in museums, it can be considered in full process. Computer technologies are increasingly capable of understanding the real world, the traditional distinction between culture and technology has become obsolete. Different approaches to creating mixed reality applications for cultural preservation and also best practices. These systems are engaging and encourage intergenerational knowledge sharing, thus have the potential to aid in the cultural preservation of partner communities (Sieck & Zaman, 2017). The growing demand for technological facilities for museums, galleries and archives has led to the need to design practical and effective solutions in this area. These facilities are intended to help with challenges such as multilingualism, to eliminate the "language barrier" and make it accessible to the people concerned (Dragoni et al., 2017).

THE INFLUENCE ON THE DIFFUSION OF NEW TECNOLOGIES

Scientific Digital Social Networks (*Redes Sociales Digitales Científicas*, RSDC) have been an important advance in the dissemination of knowledge in the scientific community, where more and more researchers agree on the need to be added to some RSDC due to the significant advantages of visibility and impact that they offer (Rodríguez-Fernández et al., 2018). The change and evolution constant of information requires an update of knowledge, as well as the filtering of it, is also favored by the use of these media (de Benito Crosetti, 2013).

Specific social networking platforms for experts, the exchange occurs between very different disciplines and areas, accelerates the dissemination of knowledge and enhances multidisciplinary collaborations, but it can also produce an overflow of knowledge. Therefore, the usefulness of social networking platforms, highlighting the importance of adopting new tools, strategies and methodologies designed to promote transversal or cross-disciplinary engagement (Jerome, 2013).

The analyzes in this study generate many potential lines such as communication, education, accessibility, trust, equality, inclusion or, thanks to the digitization of museums, eliminate “knowledge gaps” such as the “Generation gap”, adapting to the new media and the “physical gap” due to the impossibility of geographical displacement, mobility or the Covid-19 pandemic (Fuentes Morales,

2009) Knowledge Gaps or Knowledge Barriers are increasingly being considered, in fact, after the emergence of the global pandemic, many of them have come to light or have been accentuated.

Barriers in Knowledge Management were reduced in 2012 when the European Research Area (ERA) was created, for the sake of a supposed universal democratization of culture, not exempt in many cases of ethical conflicts. Thanks to digitalization, information and its extension through cultural connectivity networks based on research, creation, innovation and services are shared in terms of culture (Carrasco Garrido, 2012).

The benefits provided by the Museum when interacting with users through digital media and social networks, being open and accessible to all citizens, is to contribute to the growth of a "Cultural Society", eliminating "knowledge barriers" has been corroborated with the event of the World Health Pandemic, it´s a fascinating and very interesting topic to describe in this contribution.

Currently, the management of communication for an interconnected and plural society requires public relations policies for its visibility, users actively participate in the knowledge society and museum professionals must adapt to this new situation, both for the internal and external communication in the museum, with the interaction of the public in a virtual environment (Martínez Peláez, et al., 2012). This situation has repercussions for Cultural Organizations, such as archives, libraries, cultural exhibitions, as well as museums. The use of digital media and Social Networks by museums, contributes to the creation of on-line value, in terms of efficiency, novelties, blocking, complementarities, it even favors some performance measurements such as knowing the position of the museum from the point of view of visitors, followers on their social networks or the interest it generates, so museum professionals should think critically about the information that is exposed on the web or social networks, from the perspective of museum visitors before and after the visit . It not only favors researchers or the public interested in these topics, but also other institutions, curators and museum administrators, to give an example, which facilitates decision-making by being better informed or allocating of resources. The use of Social Networks as a means of communication requires an adequate strategy, because it can have negative consequences (Águila-Obra, 2013).

Normally social networks create sporadic interactions, with acquaintances or strangers, unilaterally or bilaterally, that do not manage to involve people in the contents or resources in their physical environment. Depending on the platform being used, there are enabled functions such as "recommend", which is interesting to make known to a greater number of people know about the achievements in research, publications or to promote a new temporary exhibition in the museum (Bravo-Torres et al., 2014).

In social networks, strong and weak interactions are created on-line, which favors creativity due to this diversity. Strong interactions often lead to offline interactions and improve the quality of knowledge received (Park et al., 2017).

It is true that more interest is shown in the information shared by users belonging to each individual's social network than in the recommendations or sharing of a third party (Kosonen, 2009).

It is such a current topic that no specialized literature has been found on this topic and the one that has been found has been at the informative level. Although there were many digital tools, there are many novelties, both in dissemination and communication, connecting public and private organizations and society, which is why there is a need for the development of ethical policies and practices in the use of ICTs, which is the technology of the future and of the present. For this, communication between research communities and professionals in ethics, information, informatics and ICTs is essential.

The value of digitization adds to that information and in its extension through cultural connectivity networks based on research, creation, innovation and shared services in the field of culture, making this a sustainable resource both economically and socially (Carrasco Garrido, 2012). In the dissemination, such as an article, a conference or publication, aspects such as confidentiality, data protection or copyright will be taken into account, aspects that are currently regulated. However, due to the increase in the use of new technologies, new situations arise that would be interesting to regulate as well. One of the points in where conflict may appear is in the comments of the users regarding a publication. User comments can bring more light to a publication, make inquiries, or receive congratulations, avoid interrupting in the case of an oral presentation or give immediacy when reflecting an idea avoiding its forgetfulness.

One of the factors that determine this behavior are the cognitive aspects of a person, so that the same publication is received by each individual in a different way. Cognitive aspects are those acquired by a person throughout his life, due to multiple factors such as his training, experience, the culture that surrounds him/her or the fact of living in one or another geographical location, they are complementary to training and work experience. In this way, even if two people have been classmates at university, the cognitive aspects will always be individual.

This is very favorable to have different points of view of the same information and occurs with all professionals from different subject, areas and levels, who work to carry out a common work.

But, on the other hand, inappropriate comments may appear due to the substance or form of the same, and in some cases protected by the anonymity of the name of the profile shown.

This is a complex circumstance that to date for large companies is being difficult to find the situation.

To avoid these situations, for example, there are different "semantic filters" to avoid displaying certain terms, although inappropriate comments can be made without swear words thanks to the use of a rich language that the systems do not detect. Another example is the creation of "closed groups" of people or entities, interested in the same topic, where users are identified, but dissemination is reduced to that area.

Currently, knowledge is sometimes favored by collaboration and documentation, facilitated in social networks, the creation of networks in a certain area, where users exchange and create knowledge, favoring the development of studies and research (Nascimbeni, 2014).

In these situations, ethics must be present, in order to continue with the processes of dissemination to both experts and citizens, but technical and legal issues also have an influence.

The world in 2020 has undergone a drastic change, due to the global pandemic of the Covid-19. It has affected all areas, there have been changes in regulations issued by the governments, and restrictions of health administrations, which directly affect the functioning of every public administration and service, as well as in the world of museums and research.

But there have also been changes devised to continue with daily activities, where creativity, imagination, illusion and initiative, among other things, which human beings have developed in this crisis, that have also been exceptional. Many of these contributions have been made in digital media, using the internet, social networks, digital platforms, etc, in order to be able to follow communications without the strict need of on site presence.

Citizen E-Science (eCC) or cyber science, is the term used for the participation of citizens in scientific projects, the use of information and communication technologies (ICT) has exploded in recent years due to the contribution of new scientific approaches, the increase in ICT devices, and even more so after the global pandemic due to Covid-19 (Finquelievich y Fischnaller, 2014).

Seeing that the use of technology inside and outside the classroom is increasingly used, it must be accepted that the way of learning and accessing knowledge is changing (Piquer, 2014, pp. 207-229).

EMPIRICAL STUDY

An empirical study has been carried out in two routes, on the one hand, a "case study" with interviews to experts and on the other hand, a study by questionnaires at state level, where the great use of new technologies and the possibility of improvement if workers are trained in it is detected.

The questionnaires have been developed considering into numerous articles for their elaboration (Medina, et al., 2019) and audit processes (Carmona-Osorio, et al., 2017).

This is a general study to ascertain the current situation of all the museums surveyed, since a more in-depth study would require a personalized study of each one of them, given their particular characteristics, such as size, subject matter, management or number of employees.

It´s a simple questionnaire in understanding and to carry out, in order to facilitate its execution. We have been considered the variables that we want to carried out in this study, and always bearing in mind that new technologies are a factor to be taken into account in many aspects, from registration, cataloging, but also communication between museum workers, as well as with experts or the general public, considering different ways, social networks, digital platforms or online training.

The answers to the questionnaires are evaluations on a scale of 1 to 5, with 5 being the highest or best valued assessment and 1 the worst valued, and NS/NC when for different reasons they cannot be assessed, due to being an external job or due to the peculiarities of each museum.

We work with the data obtained from the surveys and very visual graphs are made, with traffic light colors, in red NS/NC, does not provide information or would require a more detailed study, in dark green and light the best ratings, 5 and 4, in yellow the intermediate value 3, and in cold blue those but valued 1 and 2.

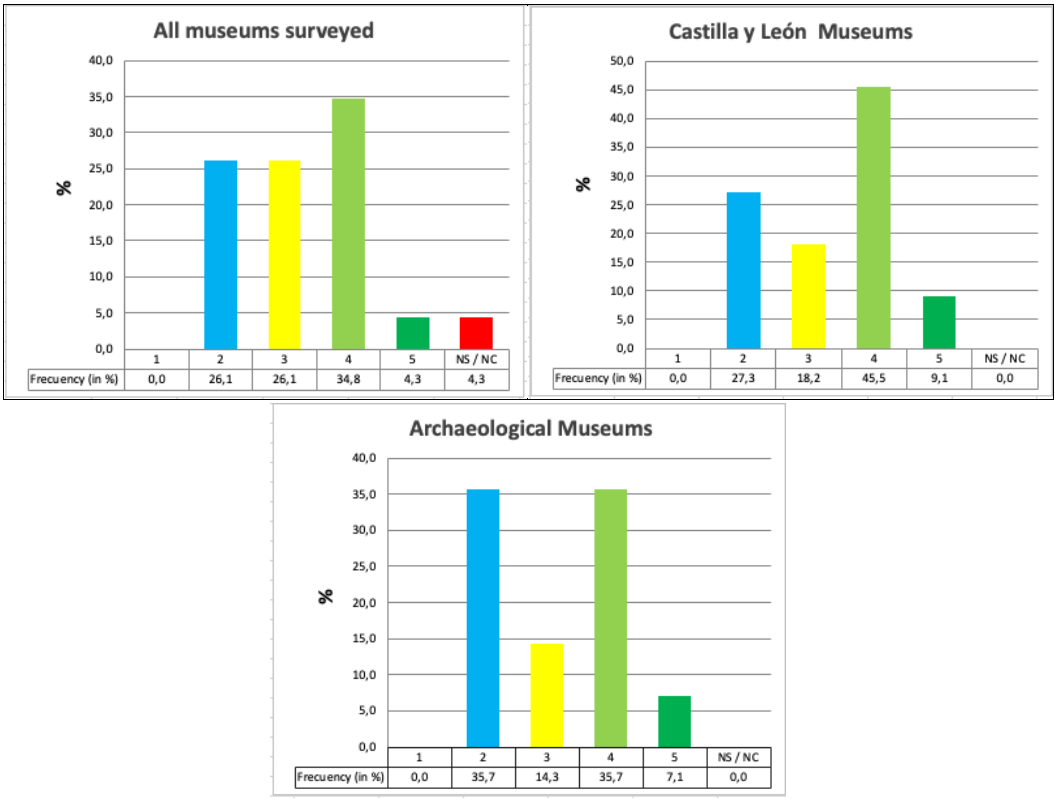
RESULTS

The results obtained from the questionnaires to experts, have been classified into three scenarios, the first is that of all the results obtained at the state level, the second at the regional level of Castilla y León and finally the museums of archaeological theme.

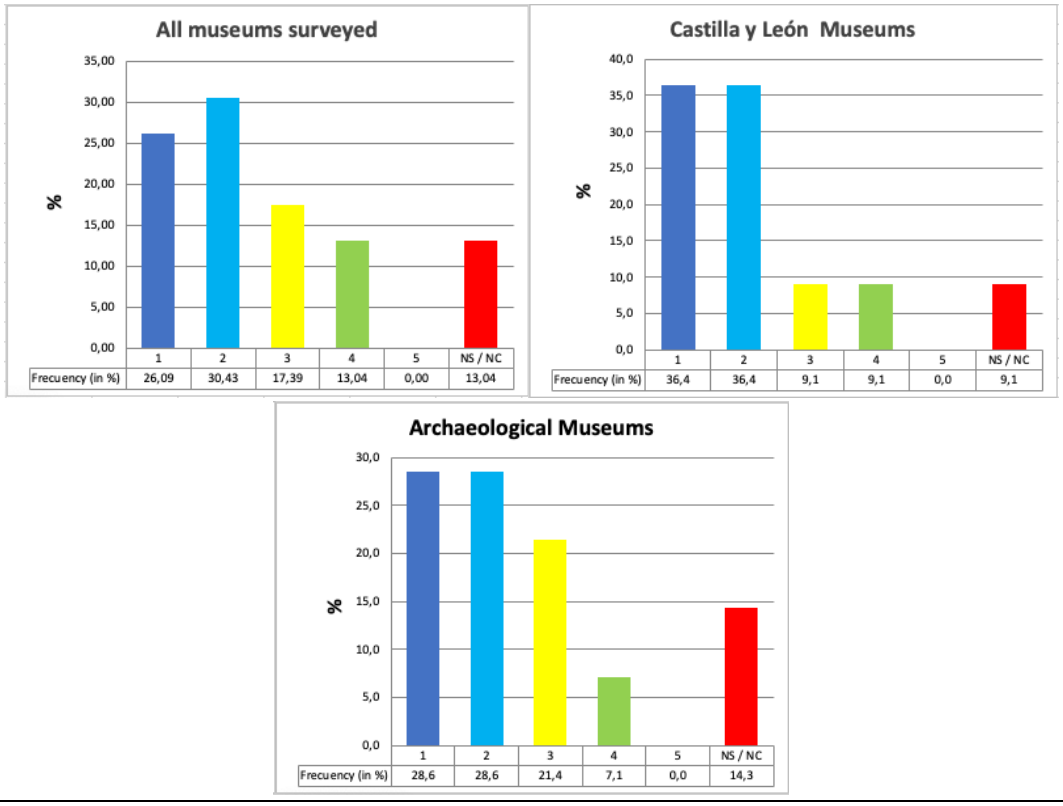
The results obtained from the questions posed are shown graphically. In the first case, the responses of museums workers perceive how the museum has elements to Share, Organize and / or Disseminate Technological Knowledge (database, web, mail, social networks, ...) and in the second case, they show their perception about the courses or training in new technologies offered by the museum to employees. Each of the two issues is shown in the three scenarios.

Figure 1. Graphs of the data obtained by the questionnaires made to experts (2020).

The workers consider that the Museum has elements to Share, Organize and / or Disseminate Technological Knowledge (database, web, mail, social networks...).



Courses on new technologies offered by the Museum workers to develop Knowledge.



Source: self-elaboration based on questionnaires (2020)

Museum workers positively perceive that they have technological means, such as databases, web or social networks to share, organize and/or disseminate knowledge in 39,1%, 54,6% y 42,8%, (All museums surveyed, Castilla y León Museums and Archaeological Museums, respectively) but they denote a lack of offer by the museum in courses of new technologies to develop its knowledge to improve the performance of its functions in 56,52%, 72.8% y 57,2% (All museums surveyed, Castilla y León Museums and Archaeological Museums, respectively).

The use of new technologies in the cultural sector is demonstrated, in line with the data obtained, that it increases the use of digital and technological training by 31%, enhances scientific dissemination by 26% and requires improvements in implementation and usability of data base algorithm by 35%.

The use of new technologies in the cultural sector is demonstrated, in line with the data obtained, that it increases the use of digital and technological training by 31%, enhances scientific dissemination by 26% and requires improvements in implementation and usability of data base algorithm by 35%.

As shown in these graphs, the museum experts surveyed perceive the current importance of the use of new technologies, but also a need for training and updating on this subject. The knowledge of a good use for communication and dissemination by these experts in digital tools and social networks, is one of the aspects that favors the performance of functions and ethics in this process.

Figure 2. The Human Evolution Museum (Burgos).

Source: self-elaboration Raquel García -Martín (2021)

At the state level, museums are indicated during the pandemic, the different measures and restrictions that must be adapted, as well as at the regional level, but the MEH also leads different initiatives to continue with its objectives in pandemic conditions.

During the time that the Museum, as well as the Atapuerca Foundation and the Research Center, with which the museum collaborates, remained closed due to the Covid-19 health alert, they continued to interact with their users and followers through all digital platforms and social networks, promote debates and scientific dissemination, book presentations and interviews and different activities, among others with different hashtags (#elMEHdesdecasa, #Quedateencasa y #CENIEHencasa). Promoted their website and content, and provided free content downloads, therefore resuming their scientific and cultural activities and encouraging users to continue getting to know the Museum virtually.

This way, the museum has been able to continue to offer "the product" to all visitors, even in a non-physical way.

The museum became a meeting place with citizens through its programming, providing science and culture encounter, albeit virtually, which favour the dissemination of knowledge to the expert and the general public, as well as contact between researchers for research collaboration and to proposing future work.

Different technological options that the museum uses to achieve one of its objectives, to bring culture, science and knowledge to the widest possible audience are described here:

1. Websites and Newspaper: Offering information of the MEH, the Archaeological Sites, the excavation campaigns and their investigations, virtual visits, activities, offering free downloads and from the dissemination point of view, various publications, educational resources, documentaries, videos, the Atapuerca Newspaper (monthly frequency, mainly in digital format, with a didactic vocation, but scientifically rigorous, it has a page in English and texts adapted for easy reading), among others.

2. App: Free application, with three routes of guided visits (MEH, Archaeological Sites and "Atapuerca Experimental Archaeology Center"), in Spanish and English. This is extremely interesting to receive explanations during the visit. It greatly favours accessibility for people with visual, hearing and mobility disabilities, and groups with functional diversity; since it has images and audios adapted to people with visual or hearing disabilities, including subtitles and video in Spanish sign language, and for those who cannot come to the museum in person.

Table 1. Annual Memories MEH. Social Networks MEH.

	2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2021 March**
TOTAL	**45.000**	**45.000**	**67.189**	**92.078**	**101.623**	**105.738**	**181.322**
Facebook	13.000	24.010	24.010	24.010	38.227	39.618	41.898
Twiter total	32.139	40.839	40.839	52.458	58.398	59.569	63.496
Twiter MEH	*17.400*	*23.800*	*23.800*	*32.158*	*34.600*	*36.320*	*38.005*
Twiter Miguelón	*9.000*	*11.300*	*11.300*	*13.500*	*15.325*	*16.200*	*16.842*
Twiter Lucy	*5.739*	*5.739*	*5.739*	*6.800*	*7.598*	*7.924*	*8.649*
Google		727	727	727			
Instagram		483	483	1.600	2.475	3.367	6.092
YouTube		939	939	1.264	2.160	2.840	5.950
Pinterest		215	215	270	305	318	361
Issuu		21	21	24	26	26	29

Source: https://www.museoevolucionhumana.com/es/memoria-2019

Table 2. Annual Memories MEH Visits MEH and Archaeological Sites.

	2014	**2015**	**2016**	**2017**	**2018**	**2019**
MEH Permanent Exhibition	151.941	147.634	148.921	150.430	150.817	151.877
National			129.691	131.128	133.451	134.319
% National			*88.92%*	*88.70%*	*88.49%*	*88.44%*

Foreign			16.155	16.711	17.366	17.558
% Foreign			*11.08%*	*11.30%*	*11.51%*	*11.56%*
TOTAL Visits MEH	**-**	**59.003**	**195.400**	**197.143**	**195.923**	**194.957**
Archaeological Sites	73.423	72.506	71.279	80.601	76.963	77.567

Source: https://www.museoevolucionhumana.com/es/memoria-2019

These tables highlight the increase over the years in both the use of technological means and the number of visits.

3. Social Networks (RRSS): The Museum has more than 180,000 followers adding all its Social Networks. (Table 1) and increase of visits (Table 2).

4. Sustainable Development Goals of the "2030 Agenda": To join the Sustainable Development goals of the 2030 Agenda, the Museum aims to develop a high quality, inclusive and equitable education program, and promote learning opportunities for all people. Promoting and supporting the ethical teaching of ICT applied to professionals (Museo de la Evolución Humana & Junta de Castilla y León, 2021). The Museum implements innovative solutions, offering educational resources and Use of streaming to broadcast face-to-face activities and workshops under the idea of "Safe culture".

The monitoring of the data that the MEH disseminates by digitization is truthful and demanding, showing its transparency portal, while it does not harm the institution and benefits. In social networks there is what the MEH wants to teach or show, what is said about it, but also the comments of users, which can sometimes be classified as anonymous and which are a representation of society, these comments have also to be ethical.

It is necessary to place in the vanguard the technological Ethical point of view (set of moral norms that govern the conduct of the person in the field of technology) in society, organizations and governments.

CONCLUSIONS

The range of possibilities offered by new technologies is very wide, it favors the accessibility of a greater number of users. The fact that it reaches a larger number of the population generates a greater number of culture seekers due to the interest generated, and the access to the general child public a "pool of culture seekers" and enhances the vocation of future researchers, but it is important Knowing the source of information to be accessed, here also appears the need to evaluate the ethics of what that source offers. Therefore, as mentioned, the collaboration of institutions, informative entities and professionals in the digital world is becoming increasingly necessary.

As a conclusion: New technologies have come to stay. Its use can facilitate Knowledge Management in all areas, which also include museums and similar meeting places for culture and science dissemination among society. A good KM, can facilitate its management and thus obtain numerous and varied benefits, help cooperation, teamwork, efficiency and leadership, and the people and research to come. It reflects the importance of the use of new technologies, common platforms that facilitate work, updating, searching and communication (García-Martín, et al., 2021) but keeping in mind the need for a Technological Ethics in its use.

ACKNOWLEDGEMENTS

We would like to thank The Human Evolution Museum for its help and accessibility, especially to General Coordinator and Head Management of Didactics.

KEYWORDS: Dissemination of knowledge, New technologies, Museum, health crisis Covid-19, Safe Culture, Generation Gap.

REFERENCES

Águila-Obra, A.P.-M. &. A.D. (2013). Web and social media usage by museums: Online value creation. *International Journal of Information Management,* 33(5), pp. 892-898.

Alonso Alcalde, R. (2018). *El Museo de la Evolución Humana MEH.* Burgos: Diario de los yacimientos de la Sierra deAtapuerca.

Bravo-Torres, J., López-Nores, M., Blanco-Fernández, Y. & Pazos-Arias, J. (2014). A Platform to Exploit Short-Lived Relationships among Mobile Users: A Case of Collective Immersive Learning. En: *Communications in Computer and Information Science.* Druskininkai,: s.n., pp. 384-395.

Carmona-Osorio, C., Ángel-Gallego, S. & Arias-Pérez, J. (2017). Strategic orientation and strategies to manage organisational knowledge and creativity ISSN: 1012-8255. *Academia Revista Latinoamericana de Administración,* 30(3), pp. 312-327.

Carrasco Garrido, R. (2012). Documentar el patrimonio : cuando la información se transforma en un recurso sostenible. *Museos.es: Revista de la Subdirección General de Museos Estatales,* 7(8), pp. 120-125.

Carrasco Garrido, R. (2012). Documentar el patrimonio : cuando la información se transforma en un recurso sostenible. *Museos.es: Revista de la Subdirección General de Museos Estatales,* 7(8), pp. 120-125.

Correa Drummond de Alvarenga Neto, R. & Gomes Vieira, J. (2011). Knowledge management at embrapa: sharing our experience on the building of a collaborative model. *Perspectivas em Gestão & Conhecimento,* 1(2), pp. 191-208.

De Benito Crosetti, B.E.A. (2013). Aggregation, Filtering and Curation for Teacher's Professional Development. *Revista de Medios y Educación,* 42, pp. 157-169.

Dragoni, M., Tonelli, S. & Moretti, G. (2017). A knowledge management architecture for digital cultural heritage. *Journal on Computing and Cultural Heritage,* 10(3), pp. 1-18.

Fernández-López, S., Rodeiro-Pazos, D., Calvo, N. & Rodríguez-Gulías, M.J. (2018). The effect of strategic knowledge management on the universities performance: an empirical approach. *Journal of Knowledge Management,* 22(3), pp. 567-586.

Finquelievich, S. y Fischnaller, C. (2014). Los Ciudadanos como "Prosumidores" de la Ciencia en la Sociedad del Conocimiento: Tendencias Mundiales con acento en América Latina. En: *Innovación abierta en la sociedad del conocimiento redes transnacionales y comunidades locales.* Buenos Aires: Instituto de Investigaciones Gino Germani, Facultad de Ciencias Sociales, UBA, pp. 64-97.

Fuentes Morales, B.A. (2009). *La Gestión de Conocimiento en las Relaciones Académico- Empresariales. Un Nuevo Enfoque para Analizar el Impacto del Conocimiento Académico.* Universidad Politécnica de Valencia ed. S.l.:Organización de Empresas, Economía Financiera y Contabilidad.

Jerome, L.W. (2013). Innovation in social networks: Knowledge spillover is not enough. *Knowledge Management Research & Practice,* 11(4), pp. 422-431.

Kosonen, M. (2009). Knowledge sharing in virtual communities – A review of the empirical research. *International Journal of Web Based Communities,* 5(2), pp. 144-165.

Martínez Peláez, A., Oliva Marañón, C. & Rodríguez Rivas, A., (2012). Public Interaction in a Virtual Platform Internal and External Communications at the Reina Sofía Museum of Art. *Telos: Cuadernos de comunicación e innovación,* Issue 90, pp. 71-78.

Medina, E., Rivera, D., El Assafiri, Y. & Medina, A. (2019). Propuesta de un cuestionario para el desarrollo de la auditoría de Gestión del Conocimiento. *ResearchGate. Universidad y Sociedad.,* 11(4), pp. 61-71.

Museo_de_la_Evolución_Humana & Junta_de_Castilla_y_León (2021). *Educación de Calidad.* S.l.:s.n.

Nascimbeni, F. (2014). La creación colaborativa del conocimiento en redes de desarrollo: lecciones aprendidas de un programa transnacional. En: *Innovación abierta en la sociedad del conocimiento: redes transnacionales y comunidades locales.* Buenos Aires: Instituto de Investigaciones Gino Germani, Facultad de Ciencias Sociales, UBA, pp. 98-119.

Park, J.Y., Im, I. & Sung, C.-S., (2017). Is social networking a waste of time? The impact of social network and knowledge characteristics on job performance. *Knowledge Management Research & Practice,* 15(4), pp. 560-571.

Piquer, M.P. (2014). Los medios de comunicación y tecnológicos como ejes de canalización y gestión del conocimiento.. *Educar,* pp. 207-229.

Rodríguez-Fernández, M., Sánchez Amboage, E., Martínez-Fernández, V. (2018). Use, knowledge and assessment of the scientific digital social networks in the Galician universities. *El profesional de la información,* 27(5), pp. 1097-1107.

Sieck, J. y Zaman, T., (2017). Closing the distance: Mixed and augmented reality, tangibles and indigenous culture preservation. En: Proceedings of the Ninth International Conference on Information and Communication Technologies and Development. S.l.:s.n., pp. 1-5

Oficial webs

Museo de la Evolución Humana de Burgos (MEH) https://www.museoevolucionhumana.com/

Fundación Atapuerca https://www.atapuerca.org/.

Centro Nacional de Investigación sobre la Evolución Humana (CENIEH) https://www.cenieh.es/.

Fundación Siglo para el Turismo y las Artes de la Junta de Castilla y León http://www.fundacionsiglo.es/web/es/fundacion-siglo.html

Junta de Castilla y León https://museoscastillayleon.jcyl.es/web/es/museos-castilla-leon.html.

Ministerio de Cultura y Deporte http://www.culturaydeporte.gob.es/cultura/museos/portada.html.

Observatorio de Museos de España (OME), Ministerio de Educación y Deporte, http://www.culturaydeporte.gob.es/observatorio-museos-espana/el-observatorio-de-museos-de-espana.html.

Consejo Internacional de Museos (ICOM) https://icom.museum/es/.

BARRIERS TO HUMANITIES AND SOCIAL SCIENCE FACULTY SUPPORTING RESPONSIBLE COMPUTING IN COMPUTING COURSES

Colleen Greer, Marty J. Wolf

Bemidji State University (USA)

Colleen.Greer@bemidjistate.edu; Marty.Wolf@bemidjistate.edu

ABSTRACT

In this paper we analyze challenges associated with infusing responsible computing discussions into humanities and social science instruction and programmatic considerations. We examine the level of leadership and intentionality necessary to support these discussions at public teaching universities in the United States. We interviewed department chairs in humanities and social science departments and surveyed faculty in those departments. Our inquiry focused on the role that computing and collaboration play in the pedagogy, curriculum, and research within departments. We asked about perceived barriers to collaboration and incorporating computing concerns into the curriculum and the role that computing and collaboration play in the tenure and promotion process. Three themes emerged from our qualitative review of the responses: digital natives and digital immigrants, crossing the divide, and tension. We conclude that the faculty participants are more focused on ethical issues surrounding delivery of courses rather than the broader issues surrounding computing.

INTRODUCTION

Discussions and scholarship regarding responsible computing and cross-disciplinary interests in associated issues of social responsibility always occur in particular timeframes and contexts. Recent global attention on issues such as privacy and facial recognition have thrown a spotlight on this discussion and affected the focus, both for those directly concerned with computing and those that find themselves directly impacted by university policy and practice surrounding computing. During the last year we have seen discussions surrounding online delivery explode as the pandemic took hold. Under COVID-19 conditions universities across in the United State reduced or eliminated face-to-face instruction and began a dramatic push to move courses to remote delivery. While there have been extensive discussions of the challenges associated with remote delivery within universities and in the press, universities, interested in keeping the doors open, have primarily focused on delivery rather than on programmatic interests and overall academic planning. Certainly, within this new, challenging environment, the emphasis has more frequently been on “how to delivery remotely” rather than how to address the structure of social issues associated with the development and use of computing technologies. The very nature of the expansion of the use of remote delivery brings into even sharper focus the issue of responsible computing, what it means, and how and when it will be embraced.

Exploring the extent to which humanities and social science faculty infuse responsible computing into instruction and program curriculum discussions is, therefore, a timely conversation. A key issue here is whether and how this incorporation has occurred, and whether it has involved cross disciplinary collaboration that highlights central issues. This research project examines the understanding of responsible computing by faculty in humanities and social science disciplines at public teaching-oriented universities in the United States and seeks to provide greater insight into the incorporation into the curriculum of responsible computing topics and the feasibility of enhancing ethical and social

considerations in computing. To understand these developments, we interviewed department chairs and surveyed faculty (pre-pandemic) regarding existing practices and interests. The results point to the challenge of understanding and interpretating responsible computing, the capacities of faculty and programs, and the attendant policies of universities.

LITERATURE REVIEW

The increased challenges of university program delivery under COVID are built upon the existing neoliberal pressures that universities have been facing since the late 1970s and the expectation that faculty will work with university administrations to maximize productivity. Faculty and programs functioning under neoliberalism experience a marked acceleration of expected efficiency while encountering a substantial increase in expected documentation of success. So, there is a new and ever-expanding focus on control and oversight (e.g., course assessments, program reviews, faculty reviews, course capacities) even as much within university systems seems to be out of control (e.g., costs, student enrollments, administrative oversight) (Giddens, 1991: 144-145; Rustin, 2016; Buroway, 2012; Collini, 2012).

Neoliberal conceptualizations have permeated all levels of university life and the university's relationship with the state. Universities, especially public universities, are now seen as profit centers, and this materializes through expectations that research will be funded by external entities, teaching will be conducted by those who are less expensive to employ, students will cover more costs, and more elements of course delivery will be automated (Feenberg 2017, Guerlac 2011). To complement these shifts in expectation, more bureaucratic systems are put in place to guide the institutions and to ensure accountability in a style reminiscent of corporate practices (Olssen and Peters 2005). The neoliberal framework has created what Buroway (2012: 139-140) calls a budgetary, regulatory, and legitimation crisis for higher education institutions, laying the groundwork for a spiral in which universities seek greater efficiency and more external forms of support. When that support comes from the business sector, the institution is commercialized in ways that erode public trust. This process in turn creates greater "public" calls for accountability from universities resulting in more bureaucratization to support auditing (Olssen and Peters 2005). This process tips the balance from a focus on instruction and research toward oversight. These processes directly impact the extent to which cross-disciplinary instruction occurs and when and how faculty expand their scholarship and instruction into other areas. In the end, significant questions remain about how knowledge is being produced and by whom.

DISCIPLINARY DISTINCTIONS IN CONTEXT

Because of neoliberal practices that encourage increasing efficiency, enhanced reliance on technology, and a reduction in cost, many universities have encouraged a shift to larger class sizes, more distance delivery (e.g., asynchronous online instruction, MOOCs, and now synchronous video), and the employment of more part-time or temporary instructors (Rhoads 2015, Guerlac 2011). The tension of this reality has impacted both those faculty already focused on computing (e.g., computer scientists) as well as those outside of computing, specifically humanities and social science faculty. Certainly, many of those in the humanities and social sciences have been exploring, through various cross-disciplinary interests, multiple facets of what digital technologies have meant for their disciplines and for the social world.

Yet challenges occur on several levels: the interpretive, the disciplinary, and the practical – implementation. At the interpretive level, the question is really whether research questions developed by humanities and social science faculty grasp the structural issues implied within the technologies

and ethics of computing. What core questions are being asked, and how is curriculum designed around those questions? Does research and teaching probe basic values and world views that inform the developments within computing or do they primarily question the resultant use and the users of technology? Certainly, these key questions address what is researched and the substance of instruction, but they also relate to the overall patterns of delivery and how to best deliver to contemporary students. Faculty are pushed to recognize the special needs of Millennial students and are encouraged to reach them in a manner that respects the fact that they are "digital natives" cognizant of what best prepares them for involvement in a digital world. These are complicated processes that are not necessarily well aligned within universities (Szrot 2015, Graff 2015/2016, Prensky 2001/2005).

This is especially the case since under neoliberalism the technical "concrete problem" has framed the intellectual space, establishing a hierarchy within the academy that privileges the technical over the human. Hence, a conversation about digital natives (i.e., students born after 1980) and digital immigrants (i.e., those born before 1980 and most faculty) is not simply one of who has adapted to what course delivery techniques, but it also encompasses a broader discussion of the structure of the university, the understandings associated with its disciplinary culture, the expectations associated with knowledge production, and who gets to deliver on what, when, and where (Prensky 2001/2005, Szrot 2015 Graff 2015, Mannheim 1936).

Prensky's (2001/2005) use of the terms digital native and digital immigrant suggests that digital immigrants, in this case the faculty outside computer science, do not typically have effective language, skills, or techniques to instruct today's students who are perceived as digital natives. As such, students are presumed to have a high level of skill using technology, to be naturally adept at synthesizing material, and to be competent at rapid learning. There is literature that raises significant questions about the viability of this understanding (Margaryan, et al. 2011, Bennett et al. 2008, Waycott et al 2010, Gallardo-Echenique et al. 2015, Šorgo et al. 2017). Yet the continuing discussion among administrators and faculty of these perceptions and the technological determinism that lies at its base affects understandings of how to invest university funds and how to structure curriculum and professional development (Selwyn 2012). They also undermine substantive conversations about the ethical and social issues of technology and computing and move the discussion into a practical delivery analysis that leaves many aspects of the impact of computing unexplored while reemphasizing elements of training for more efficient delivery.

As the scholarly conversation about digital native and digital immigrant moves directly to issues of training for faculty it reifies a divide along the lines of technical delivery. Leaving aside substantive issues associated with knowledge production and development, universities focus on how to best assist employees in their adaptation to enhanced digital experiences. Data are often gathered from faculty to determine interest and ability, and once an organization has these data, they can use information about what creates meaning for an employee to better train individuals on new computing technologies (Breen 2018, Mansbach and Austin 2018). This is a fairly common practice, and as Kesharwani (2020:14) indicates, employers who explore the characteristics of technology users in this manner can then create training programs designed around those characteristics and promote them as enhancing "self-efficacy." The process of labelling faculty as digital immigrants and creating trainings to enhance connections to identified digital natives highlights how data and interpretations of that data are used to broadly label and affect a particular end—training—rather than establish a structural review (Selwyn 2012). In other words, the push is for adaptation without always asking to what end.

Faculty in the humanities and social sciences have varied behavioral preferences, professional understandings, and social advantage that inform their experience and interest (Kesharwani 2020,

Sharafi et al. 2006, Mansbach and Austin 2018). If being a digital native means immediate, uncritical acceptance of technology use, then many faculty in the humanities and social sciences find themselves labelled as outsiders when they question remote delivery and the value of engaging students through digital means. Yet many faculty in these disciplines either are more proficient than those they are instructing, or they seek to be (Waycott et al. 2010). The digital native and immigrant conversation while significant, misses the deeper question of what it means to be a "native" and moves quickly to an attempt to address how to effectively train digital immigrants for quicker assimilation (Šorgo et al. 2017, Breen 2018, Mansbach and Austin 2018). Certainly, some faculty in the humanities have explored how best to weave together an emphasis on instruction with pedagogy that teaches students how to be proficient in digital environments (Guerlac 2011). So, focusing on faculty as digital immigrants does not adequately reflect the complex capacities of these faculty. Examining these ideas with an ethnic analysis lens (Gordon 1964) highlights the problematic nature of this conversation and opens the door to rethinking the level at which involvement and collaboration across disciplines might be leveraged.

If, for a moment, we accept the digital native and digital immigrant classification and the essentialist and assimilationist understandings that it implies, we can then note that we should be seeking to better address the structural, cultural, and interpersonal realities that align with assimilationist interpretations (Gordon 1964). To assume an essentialist viewpoint of digital native not only misses the complex structural interplay involved as people engage systems and each other, but it negates the variability of ability, interpretation, and levels of understanding within particular environments. In addition, the use of the terminology refocuses attention in a manner that suggests technological determinism and undermines deeper cross-conversation among disciplines (Selwyn 2012). This tension was evidenced by participants in our study.

The quick move to remote delivery at the onset of the pandemic brought the digital divide into sharp relief, undermining the essentialist understanding of the digital native/immigrant taxonomy. The pandemic also exacerbated tensions, both ideological and practical, surrounding the incorporation and use of various forms of technology. When faculty have a heightened concern about the digital divide and the impact that it has on outcomes for less privileged and basic-digital learners, broader conversations regarding collaboration to enhance ethical and social understandings related to computing are not well developed (Šorgo et al. 2017). Our research looks at where and when cross conversations occur, what sidelines those conversations, and how deeper and more comprehensive efforts toward cross-disciplinary collaborations might infuse greater awareness of the central ethical and social concerns associated with our digital experiences.

METHOD

Information from humanities and social science faculty on responsible computing was collected as part of a larger study designed to explore how and when faculty in computer science, humanities and social science disciplines engage in research and instruction on ethics and social issues related to computing. A mixed method design was used. It began with identifying and selecting state, public universities that had humanities/social science and computer science programs or departments. Each had between 3 and 65 faculty. The two-phased approach is based in critical methodology and designed to both interpret and engage (Babbie, 2016). Institutional Review Board approval was received for all instruments and processes used in this research.

Phase one of the study was a qualitative interview with identified chairs. Qualitative interviews were conducted via an audio-only Zoom meeting, and informed consent and debriefing documents were provided to all interviewees as part of the process. Questions were created to identify when and how

computing concerns were incorporated into the humanities and social science curriculum. Separate questions were asked of the computer science faculty to discern whether ethnics and social responsibility is being infused into the computer science curriculum. Our findings from the computer science interviews and surveys are discussed in (Greer & Wolf, 2020). The overall intent was to examine disciplinary distinctions so that a better understanding of each discipline's approach will inform cross-collaborative instruction and research. To ascertain the extent to which collaboration occurs or could occur, we asked about levels of collaboration related to instruction and scholarship, and the extent to which responsible computing was aligned with strategic or master academic planning at their institution. In all interviews four basic questions were asked, with additional probes used to ensure clarity.

Phase two of the study was a survey of faculty within the departments where department chairs specified interest. A short survey instrument was distributed to all faculty in those departments. Seventeen survey questions were created and a Qualtrics survey instrument was used to distribute the questions to humanities/social science faculty. (Questions used are available upon request.) In addition to questions related to how important responsible computing is to instruction and research in the humanities and social sciences, we asked the extent to which responsible computing occurred across the levels of their discipline and the degree to which they were satisfied with their knowledge related to responsible computing. There were also basic demographic questions.

We analyzed the chair interview transcripts and the survey results from the faculty. Of the 78 surveys distributed, 25 have been completed. The analysis focused on a qualitative interpretation of the data collected. Qualitative analyses provided an opportunity for deeper interpretation of social worlds and the impact of particular interactions (Berg 2009). Interview transcripts were coded using an open and focused coding process and responses were then examined in light of survey results (Berg 2009, Katz 1983). Themes of digital native/digital immigrant, crossing the divide, and tension emerged from a review of the coded transcripts. Descriptive information from the surveys will be used to support and clarify data found within the interviews.

ANALYSIS

Digital Immigrant, Digital Native

Our discussions with department chairs highlighted the challenges associated with cultural divisions of the digital world. One chair noted a faculty member who "was giving the students a lot of time to complete the quiz, and so the students constructed a private Facebook group where they were sharing the answers to the quiz." C1 Addressing these sorts of cultural issues, they noted that faculty in their areas were . . . "slow on the uptake. I don't know if it's the generational gap. . ." C1, and that, "faculty knowledge about that, that is probably one of the biggest barriers. Particularly since we've hired mostly faculty members who do qualitative research." C2 These comments point to the sense of digital native, digital immigrant present among faculty in the humanities and social science disciplines and the perceived difficulty of bringing everyone to an awareness that digital communication and computing is, while daunting, a reality. Building on interpretations by Prensky (2001/2005) we see a sense of the divide and an acceptance of that gap as a "barrier." In addition, the assertion that the gap may also be related to a distinction in research type (i.e., qualitative versus quantitative) suggests that digital understandings primarily align with computing technologies and quantitative approaches while qualitative research is divorced from digital understandings and use. We see here a labelling of faculty, an acceptance of technological determinism, and a reification of ideological interpretations and practices associated with this understanding.

The concepts of digital native and digital immigrants were used by most interviewees, and frequently the discussion moved between a discussion of distinctions based on age, as noted above, to an assertion of progress. Age was cited both through reference to a "new age revolution" and to doctoral experiences that did not contain digital approaches as part of the education to specialization. At points it was expressed as it is "… not a person's specialty." However, this was typically in reference to a subfield in a particular discipline, for example, when a faculty member's focus was not on digital humanities or digital studies. Often these subfields were directly associated with better understandings of computing and greater awareness of data literacy. Yet there was no direct evidence that the subfield interpretations involved significant understandings of the ethical issues associated with computing, but rather that the ability to engage in substantive use of information technologies and data was enhanced.

The sense of an assimilationist agenda within this conversation is evident. As one department chair stated, "Look, if you want to teach in the summer, you better learn to love teaching online and learn how to do it because that's the only way these courses are going to make." C2 This speaks to adaptations faculty must make as part of the expectations of incorporation into the next iteration of university life.

> "I became chair, this is my sixth year, and when I became chair we offered virtually no online courses … but there was a push from the university to move toward that … it took off in 14, 15 … We had a summer online program … watched it go from entirely face-to-face to practically, entirely online." C2
>
> "… there's a good bit of talk, we have a brand new president at the university, about incorporating different aspects of computing into our classes at a variety of levels and a variety of classes, …" C4
>
> "Some people have thought a good deal about it, some people not much at all. So trying to get it to find its way into the curriculum in some formal way just simply involves a good bit of work." C4
>
> "… our junior faculty has to be thinking about these things because as soon as they come in they have to go through some of the trainings which involve working with Canvas, working with online … And all of the rest of us, there's less of an onus on us." C5
>
> "And I mean they [students] really think that it's a part of our job now to be on Canvas, which is always kind of surprising to me. But in any case, that's the expectation." C5
>
> "… the faculty Senate group that's grown up in the past few years I think is partly meaning to respond to what that office is doing … the office of teaching and learning is run by faculty, but it's also kind of just like administrative initiative …" C5
>
> "And so, the way that online teaching is always pitched or hybrid teaching is always pitched by the administration is that, it's more inclusive the idea is … that students surveys say, or suggest that students who have mixed schedule, meaning some face-to-face classes and some either hybrid or online that they're much more likely to be retained …" C5

Paralleling, in interesting ways, the language of cultural and structural incorporation, we can see here how it is not only the presence of a new technology that informs relationships and university policy and practice, but the sense of the interactive aspects of that environment and the expectations of normalization and conformity that follow.

Milton Gordon (1964), in his seminal work on assimilation in American life, points to the cultural, structural, and interpersonal patterns that accompany the arrival and entrance of new migrants into the society. In a similar vein we see, contained within the quotes above, a cultural, structural, and relational understanding. Culturally, there is an expectation that faculty will achieve various levels of awareness and acceptance, certainly at the level of incorporating various forms of digital delivery into classes. Managing discussion post options or online chats in learning management systems, speaking the language of what it means to connect with students, and proficiency in the pedagogical approaches of that connection are all expected. However, these quotes go farther and suggest that in addition to following the current fad of presentation and performance, there is an expectation of shifting values and accepting the face-level values and associated ethics embedded in computing technology and remote delivery. Structurally, the incorporation of remote delivery suggests an embedding of the technology and the faculty person in the institution in a manner that redevelops the role of faculty and the institutional arrangements while it redirects the institution in terms of its strategic plan, university initiatives, and finances. Interpersonally, the normalization and conformity carries with it an expectation of faculty linking with staff in technology services who support platform delivery mechanisms, with students to help them, and in an interdisciplinary manner with other faculty who are "better equipped" for remote delivery. In all of this, however, there is often an absence of discussion about computing, ethics, substantive realities around relationship development, and driving neoliberal interests of individualism and profit.

Yet as department chairs discussed engagement with digital technologies they pointed to remote delivery, adherence to human subjects reviews, and use of various forms of software to engage in data analysis. As one department chair put it, the faculty are:

> "Trying to even teach our students how to use computers appropriately for research. I think that's a big task that we have right now … the two methods courses, they have had to evolve to deal with technological changes or how to find resources or how to find sources, and also data analysis." C3

In addition, the descriptive statistics from the surveys show that many faculty in the humanities and social sciences see responsible computing as important to instruction and to research. Of the 20 responses to the survey question on this topic, 85% saw responsible computing as important or somewhat important to their instruction and 65% saw responsible computing as important or somewhat important to their research. Irrespective of their initial training or their particular methodological approach, the merits of including social issues and ethics associated with computing were seen as significant to their work. The need to cross the divide was evident from their responses, yet there are questions of how and when to cross the divide and what mechanisms allow the bridging to occur.

Crossing the Divide

Moving beyond disciplinary specialization requires crossing into areas of expertise that have often gone unexplored, or which represent a sense of risk, risk in terms of whether it is possible to serve as expert, risk of encroaching on another's area of expertise, and risk of time spent without reward. For faculty in the humanities and social sciences there are additional risks to perception and to status. Having their abilities and understandings exposed to the review of those in technology areas whose

territory they are now exploring, these faculty risk being accused of encroachment into areas considered the territory of other experts and risk undermining their own discipline, delivery, and programming through a perception of having accepted the technological determinism of the digital without reciprocation. Finally for these faculty, there is a question of whether I, as a faculty person from a discipline outside computing, receive status and recognition for responsible computing related work that goes beyond my field. Examining responses from department chairs, it is in understandings of interdisciplinarity and collaboration that we see various elements of connectivity and related challenges. Key here are what level of connection is possible, among whom, for what purpose, and if in scholarship, what type. As department chairs pointed out faculty in the humanities and the social sciences engage in instruction and research on various aspects of ethics and social issues associated with computing, yet the chairs' commentary shows that the engagement with these topics is not consistent across all the faculty and varies based on areas of interest and perceived forms of support.

> "... I mean I can think of one person who is working on research or scholarship that has to do with computing issues and thinking through issues of AI and how that deals with humanism and post humanism ... I can think of another philosopher in my department who's working on games and gaming ... though he's really more interested in aesthetics than in the technology aspect of games per se ..." C5
>
> "So, some will pick up digital ethics, or some toward things related to computing … I couldn't say I could give you a good number." C4
>
> "I have probably two faculty members that look into ... well more broadly speaking kind of more social media with computing and trying to understand big data." C3

These quotes demonstrate that faculty in the humanities and social sciences are working with or toward responsible computing issues yet trying to retain central connections with their own disciplinary approaches and understandings. Having a foot in both worlds is a challenging endeavor, and attempting to identify what it means to incorporate, beyond delivery techniques, topics associated with responsible computing in the classroom means deep assessment of what must be delivered, when, and how.

Considering what it means to systematically engage in curricular change around topics of computing ethics and social responsibility, one department chair noted:

> "I think we're slowly starting to think about that ... I don't know if it's just my faculty or not, but we're kind of slow on the uptake." C3

And another indicated:

> "Some people have thought a good deal about it, some people not much at all. So trying to get it to find its way into the curriculum in some formal way just simply involves a good bit of work … it would probably be a kind of sub-specialty that you were picking up. There isn't anybody working directly in it." C4

At points, as the department chairs discussed potential curricular changes that would cross into the areas of responsible computing, they focused in on one course, often a general education course on ethics or values, and how that could be changed to create deeper student awareness. One chair stated:

> "... we thought about … dividing the class up, not necessarily by philosophers, but by key life issues … and I think one of the most provocative for students would be computing and everything that's involved with the technological age in which we live … a big shift from the way the class has often been taught …" C5

Yet, it is not that faculty in these disciplinary areas are largely avoiding the issue. Faculty survey responses on the question of delivery of responsible computing across the levels of instruction shows that with third-year students faculty engage to a great or some extent 70% of the time. This shows important cross-over of interest and approach. During the first and second year the percentage of time drops to approximately 58%, and in the fourth year it once again declines. These percentages demonstrate not only the sense that engaging students on these topics is significant, but the perceived placement of where it should be addressed across the degree process.

Many faculty have participated in joint scholarship with faculty in other disciplines and have explored topics with and for students that take them beyond their disciplinary home. However, finding cross-disciplinary or interdisciplinary scholarship is much more frequent than doing so with instruction. This is often due to the administrative barriers that exist within university settings. Identifying ways to give credit to two instructors for team teaching or structuring FTE in a manner that encourages intermittent sharing of knowledge from another discipline seems out of reach for compensation and accounting purposes. This leaves faculty with little recourse, even as they seek creative options for interdisciplinary delivery.

Tensions

While some faculty in the humanities and social sciences are already crossing the divide and others are exploring how to make their way into responsible computing analyses and instruction, tensions around what it means to move beyond disciplinary boundaries are persistent. Some of the tensions exemplify the extent to which administrators, conscious of neoliberal goals and efficiency needs, are pushing for greater online delivery. Other tensions surround the extent to which the digital native/digital immigrant imagery has legitimacy, or whether the imagery hides a much more complex reality in which Millennial students are ill-prepared, beyond basic application skills, to understand data and to engage in critical analyses. Finally, the push to remote learning through various educational platforms siphons attention away from more substantive discussions of university purpose, intellectual development, and collaborative options that address ethical issues of computing and that allow faculty to envision the future.

As one department chair indicated, administrative initiatives for using digital platforms are often supported by segments of faculty leadership raising challenges across faculty contingents:

> "… it's more coming from the faculty in that faculty Senate group where you have the rest of us saying, whoa, are there some things we should be considering in terms of how instruction is being handled online, how the materials are conveyed and what are the big questions that

> students should be thinking about [it] in terms of their intellectual labor and how that interacts with the digital platforms and the digital. … It's a culture that we all live in." C5
>
> "We've got some older faculty, they aren't interested in that kind of thing." C3

Even for departments making the move online

> "there have been questions about what about the integrity of the tests. How do I know that my test is secure? How do I know that students aren't cheating?" C2

Yet,

> "maybe there are some best practices that can be involved in terms of encouraging students to be more ethical and responsible with the use of the internet, but instructor best practice as well in terms of trying to ensure academic integrity and also maybe even worrying about issues of privacy and surveillance." C1

The focus on efficiencies also undermines attempts at collaborative instruction efforts. While some faculty are interested, they find it difficult to identify mechanisms to make it work and maintain the expected efficiencies and their programs. As one department chair noted,

> "… collaboration in the classroom can look more expensive in terms of FTE to administrators and therefore it makes it, for instance, more difficult for me to sometimes want to support it even though ideologically I want to support it … it makes me nervous that then if I'm asking, for instance for another faculty line that the administration would say something like, well you have people co-teaching classes, so why wouldn't you just have them teach their own classes and then you wouldn't need another faculty member." C5

They went on to note how IT infrastructure can discourage collaboration,

> "... the current software we use makes it difficult to do [cross lists]. And so, they you have software driving curriculum, which is of course something we're always pushing back against." C5

The tensions are also evident as faculty attempt to identify how best to instruct Millennial and Generation Z students and whether or when these needs generate cross-disciplinary interests in responsible computing:

> "We find that Millennials and as well as the newest generation coming in, Generation Z, hears about technology, have this supercomputer in their hand, but they don't know how to use it

> appropriately when it comes to collecting data or collecting research or even using it to find out information." C3

The challenges of working with administrative initiatives to move to enhanced remote delivery, electronic portfolios, or other digital assessment tools create tension within and among faculty. Some faculty who are focused on reaching students based on information about their interests and engaging them in the new digital age embrace various digital forms of communication and discussion. Others, worried about substantive questions of student literacy and critical ability, ethical issues associated with computing, or their own abilities related to the use of digital tools raise warning flags and resist. Yet the online instruction platforms and recent experiences during the COVID-19 pandemic are forcing faculty to identify how to move and then move forward. This is supported by faculty responses to the survey question regarding what curricular changes could be made to incorporate computer ethics and social responsibility:

> "All of my courses include this content."
>
> "I am working now to integrate discussion of research methods and sources (research and library skills) with broader discussion of computer information, web searching, and engagement."
>
> As department chairs consider attempts to shift to enhanced engagement on these topics, they demonstrated various levels of awareness, engagement, and tension:
>
> "You mean in terms of like having instructors incorporated into any number of classes or having a specific class in the curriculum? I haven't really thought about that before to be honest, but as I've referenced before we're all becoming more and more dependent on these things and more and more instructors in my department are using online platforms like Blackboard. … And it's occurring to me now that I'm speaking to you that perhaps … some sort of statement or discussion or lecture or something having to do with issues pertaining to the ethics and social responsibility surrounding use of that technology." C1
>
> "I don't think we've had a lot of discussion about it other than policy set down by the university as to what we should do with our computers and what we shouldn't." C3
>
> "… and one of the things is getting a digital humanities degree … trying to craft for students that, despite the fact that what we talk about most in the humanities or the arts as the objects that human beings have created, that are expressive of human culture, that certainly technologies we see in the current time are those objects too." C5

There is an awareness of the need to engage in substantive and pedagogical considerations related to responsible computing and associated social issues. Many faculty can see the privacy challenges, the connectivity and access challenges, and the need to ensure that students are well versed in the use of the digital. At the same time there are significant points of deep disconnect between the delivery platforms that are part of the focus of day-to-day life in the university, and the underlying considerations of computing ethics and social issues that are associated with their development and use. The ongoing tensions are real for faculty and are often expressed through comments about the pressures they experience from administrative initiatives, fractured time commitments, and the drive to technology as answer.

CONCLUSION

Information from this research sheds light on how humanities and social science faculty are engaging with issues related to responsible computing. While, based on existing definitions, none of the faculty could be considered digital natives, the language used to define the landscape of digital native/digital immigrant is demonstrated to not effectively communicate the complex connections that these faculty have to digital technology and connection with students. Some faculty are quite well-versed in the various forms of remote delivery and are already exploring topics relevant to computing ethics and social responsibility. Others, while less directly engaged, are recognizing the necessity of examining the underlying structural and cultural issues associated with digital development and use. Many have already crossed the perceived divide, are teaching remotely, and are incorporating new topics about computing ethics into their courses to better prepare students. Yet, this incorporation does not mean that there are not ongoing tensions related to the university environments in which they instruct and engage in scholarly work. In fact, the impact of COVID-19 on university life has meant the creation of increased pressures to instruct remotely and has required more extensive use of digital platforms. The emphasis within universities on use and the obvious ethics surrounding use and privacy, directs the gaze away from other central questions of the intention of development, human relationship patterns, and how humanity can maintain empathy and care (Foucault 1977). As the gaze is refocused, it is easy to miss the control over the substance of the intellectual discussions related to digital developments as well as the use of digital platforms. As one department chair noted, faculty should be framing discussions that directly address "… the big questions that students should be thinking about …" C5 These broader issues are at the heart of necessary considerations for cross-disciplinary discussion and action.

KEYWORDS: computing ethics, teaching computing ethics, collaboration in higher education, integrating computing ethics, social responsibility in computing.

REFERENCES

Babbie, E. (2016). The practice of social research (14th ed.). Boston, MA: Cengage Learning.

Bennett, S., Maton, K. & Kervin, L. (2008). The 'digital natives' debate: A critical review of the evidence. *British Journal of Educational Technology*, 39(5), 775-786.

Berg, B. (2009). Qualitative research methods for the social sciences. Boston: Allyn & Bacon.

Breen, P. (2018). Developing educators for the digital age: A framework for capturing knowledge in action. London: University of Westminster Press.

Buroway, M. (2012). The great American university. Contemporary Sociology, 41(2), 139-149.

Collini, S. (2012). What are universities for? London: Penguin Books.

Feenberg, A. (2017). The online education controversy and the future of the university. *Foundations of Science*, 22, 363-371.

Foucault, M. (1977). Discipline and punishment: The birth of the prison. New York, NY: Vintage.

Gallardo-Echenique, E.E., Marqués-Molías, L., Bullen, M., Strijbos, J. (2015). Let's talk about digital learners in the digital era. *International Review of Research on Open and Distributed Learning,* 16(3), 156-188.

Giddens, A. (1991). Modernity and self-identity: Self and society in the late modern age. Stanford: Stanford University Press.

Gordon, M.M. (1964). Assimilation in American life: The role of race, religion, and national origins. New York, NY: Oxford University Press.

Graff, H.J. (2015). Undisciplining knowledge: Interdisciplinarity in the Twentieth Century. Baltimore: John Hopkins University Press.

Graff, H. (2016). The "problem" of interdisciplinarity in theory, practice and history. *Social Science History*, 40(4), 775-803.

Greer, C. & Wolf, M.J. (June 2020) "Overcoming barriers to including ethics and social responsibility in computing courses" ETHICOMP 2020.

Guerlac, S. (2011). Humanities 2.0: E-learning in the digital world. Representations, 116(1), 102-127.

Katz, J. (1983). A theory of qualitative methodology: The social system of analytic fieldwork. In Robert M. Emerson (Ed.), Contemporary field research: A collection of readings (pp. 127-148). Prospect Heights, IL: Waveland.

Kesharwani, A. (2020). Do (how) digital natives adopt a new technology differently than digital immigrants? A longitudinal study. *Information and Managemen*t, 57.

Mannheim, Karl. (1936). Ideology and utopia. New York, NY: Harcourt, Brace and World, Inc.

Mansbach J. & Austin, A.E. (2018). Nuanced perspectives about online teaching: Mid-career and senior faculty voices reflecting on academic work in the digital age. *Innovative Higher Education*, 43, 257-272.

Margaryan, A., Littlejohn, A. & Vojt, G. (2011). Are digital natives a myth or a reality? University students' use of digital technologies. *Computers & Education*, 56, 429-440.

Olssen, M. & Peters, M.A. (2005). Neoliberalism, higher education and the knowledge economy: from the free market to knowledge capitalism. *Journal of Education Policy,* 20(3), 313-345.

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon*, 9(5).

Prensky, M. (2005). Listen to the natives. *Educational Leadership*, 63(4), 8-13.

Rhoads, R.A., Camacho, M.S., Toven-Lindsey, B., & Lozano, J.B. (2015). The massive open online course movement, xMOOCs and faculty labor. *The Review of Higher Education*, 38(3), 397-424.

Rustin, M. (2016). The neoliberal university and its alternatives. Soundings, 63, 147-170.

Selwyn, N. (2012). Making sense of young people, education and digital technology: The role of sociological theory. Oxford Review of Education, 38(1), 81-96.

Sharafi, P., Hedman, L. & Montgomery, H. (2006). Using information technology: engagement modes, flow experience and personality orientations. *Computers in Human Behavior,* 22(5), 899-916.

Šorgo, A., Bortol, T., Dolničar, D. & Podgornik, B.B. (2017). Attributes of digital natives as predictors of information literacy in higher education. *British Journal of Educational Technology*, 48(3), 749-767.

Szrot, Lukas. (2015). The idols of modernity: The humanity of science and the science of humanity. [Master's Thesis, The University of Texas at Arlington].

Waycott, J., Bennett, S., Kennedy, G, Dalgarno, B., Gray, K. 2010. Digital divides? Student and staff perceptions of information and communication technologies. *Computers & Education*, 54, 1202-1211.

A REVIEW OF TRAFFIC ANALYSIS ATTACKS AND COUNTERMEASURES IN MOBILE AGENTS' NETWORKS

Rafał Leszczyna

Gdańsk University of Technology, Faculty of Management and Economics (Poland)

rle@zie.pg.edu.pl

ABSTRACT

For traditional, message-based communication, traffic analysis has been already studied for over three decades and during that time various attacks have been recognised. As far as mobile agents' networks are concerned only a few, specific-scope studies have been conducted. This leaves a gap that needs to be addressed as nowadays, in the era of Big Data, the Internet of Things, Smart Infrastructures and growing concerns for privacy, the subject gains particular importance. This paper presents the results of a literature study that aimed at identifying traffic analysing attacks and countermeasures in mobile agents' networks. No limiting assumptions are made in regard to the complexity or size of agent networks. Also, various types of attackers' configurations have been analysed and referred to each attack. The results enable building appropriate threat models for cybersecurity management or when designing new security solutions.

INTRODUCTION

Mobile agents are not a new concept. They have been studied for more than two decades. However, nowadays, in the era of Big Data, the Internet of Things (IoT), Smart (Cities, Grids, Factories, Buildings, etc.), Sensor Networks or Vehicular Networks, they regain attention due to the new applications, for which they appear to be particularly suitable (Calvaresi et al., 2019; Kampik et al., 2019; Kem & Ksontini, 2019; Urra et al., 2017; Yang et al., 2017; Zrari et al., 2015). For instance, business process management in the IoT environments (Kampik et al., 2019) or smart grid energy management systems (Kem & Ksontini, 2019) can be implemented using the agents technology. Coupling multi-agent systems with chatbots in the context of health-support programs have been also studied recently (Calvaresi et al., 2019). Another example is vehicular networks, where the exchange of sensor data between automobiles by using short-range radio signals exposes certain challenges that can be effectively addressed by mobile agents (Urra et al., 2017).

Mobile agents autonomously roam over networks to perform tasks on behalf of users (Mayowa et al., 2016; Yang et al., 2017). Technological benefits associated with the agents include bandwidth conservation, reduction of total completion time, latencies reduction, disconnected operation and mobile computing, load balancing, dynamic deployment and improved querying of various information sources (Gray et al., 2000). The challenges include lower performance in certain application areas, a limited number of mature development-oriented platforms or higher costs of implementation (Gray et al., 2002; Isern & Moreno, 2016; Luck et al., 2003; Mayowa et al., 2016; Yang et al., 2017; Zrari et al., 2015). The indubitable difficulty for mobile agents is to assure their security. This is especially because, during migration, an agent is dependent on the hosts it passes. During the years of research, various attacks have been identified and countermeasures proposed (Bouchemal & Maamri, 2016; Jolly & Batra, 2019; Madkour et al., 2014; Sanae et al., 2019). A separate subject that requires particular attention today, when Internet users' privacy has become a great concern (Choi et al., 2018;

Lopez et al., 2017), is to assure that in sensitive application areas, such as healthcare, insurance, banking and many others (Isern & Moreno, 2016; Pellungrini et al., 2017; Xia et al., 2017) agents could not be traced to their owners, neither by reading the agents' data nor by performing traffic analysis.

The term *traffic analysis* (TA) comes from military intelligence. It describes the process of "intercepting and examining messages to deduce information from their patterns of communication" (Sobh & Elleithy, 2015). With the advent of the Internet, TA was soon applied to Information Technology. While the subject of traffic analysis has been studied extensively for classical, message-based communication (Cottrell, 1995; Dolev & Ostrobsky, 2000; Goldberg & Wagner, 1998; Gülcü & Tsudik, 1996; Kesdogan et al., 1998; Raymond, 2001; Reiter & Rubin, 1998), only a few studies have been dedicated to mobile agent networks. Kulesza et al. (Kulesza et al., n.d., 2006) formulate the problem of traffic analysis for mobile agents, outline the elements of the security model, with selected adversaries (listening adversary and active adversary), assumptions, threats (e.g. following agents, tracing an agent's return route, observing network nodes or side-channel attacks) and potential countermeasures. The authors focus on complex environments with large numbers of mobile agents roaming autonomously that resemble wireless networks. In this way, they can adopt the approach of Matt Blaze et al. (Blaze et al., 2005) to dedicated to traffic analysis in wireless networks. Several studies have been dedicated to the reverse subject, namely the mobile agents used for traffic analysis. For instance, Dasgupta and Brian (Dasgupta & Brian, 2001) apply this approach to build a distributed intrusion detection system that monitors network traffic and emulates mechanisms of a natural immune system using mobile agents.

This paper presents the results of a broad study of the literature dedicated to the traffic analysis in mobile agent systems and classical message-based communication. No limiting assumptions are made about the complexity or size of agent networks. Conversely – all types are taken into consideration i.e. the large, and crowded environments, but also small, and deserted ones, where only a few agents roam. In the following sections, the traffic analysis adversaries are described, followed by the presentation of attacks and potential countermeasures. For each attack, the required setting of an adversary (the adversary type) is indicated. Also, the countermeasures have been identified based on a broad study of the available literature. The paper ends with concluding remarks.

ADVERSARIES

The literature on traffic analysis (Dolev & Ostrobsky, 2000; Lindell, 2010; Raymond, 2001; Syverson et al., 2000) distinguishes the following types of attackers:

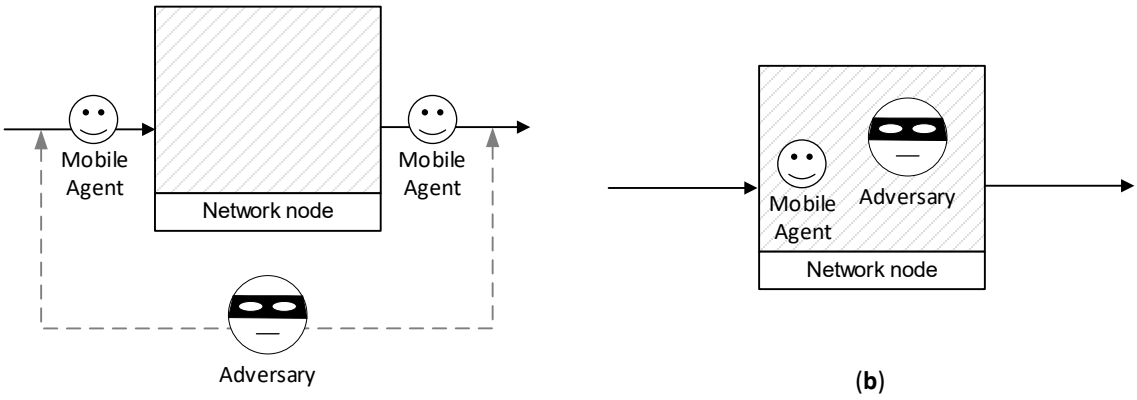
- *Internal/external* – The *internal adversary* (see Figure 1b) is the adversary who succeeded in compromising a network node (e.g. a host, or a network device that uses an operating system), and can observe agents within it (*passive* – see below) or even may obtain control over the node (*active* – see below). The external adversary (see Figure 1a) is the one who aims at a particular node but was able to compromise (for example due to the lack of security of an encrypted link) only the communication channels leading to and from it. Internal attackers are viewed as more potent than external attackers since they have access to more resources and in particular, they may observe the agent's behaviour at the node to distinguish the agent from others. Considering a particular node, an internal attacker who managed to compromise the node has at least as a good view as the external attackers who managed to compromise all channels entering and leaving the node. On the other side, the external attackers can observe agents coming from/to other nodes if only they pass through the channel the adversary observes. Note that the distinction between the two types of attackers must be made in the context of a particular network node. Thus, an attacker who compromised several nodes but

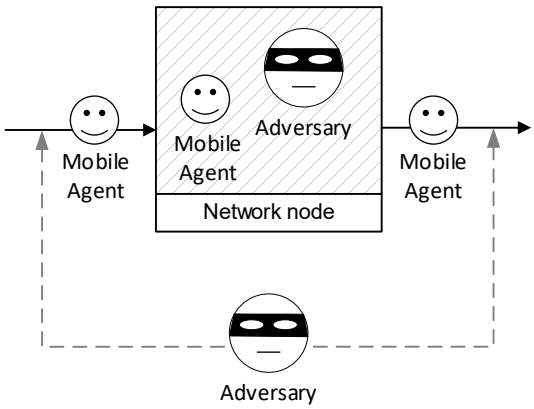
is aiming at another one (which they were not able to compromise) should be viewed as external from the point of view of the targeted node (Raymond, 2001).

- *Omnipresent / k-listening* – The adversary may succeed in attacking all nodes (then the adversary is called the *omnipresent* adversary), or *k* of them (the adversary is called the *k-listening* adversary (Dolev & Ostrobsky, 2000)). In particular, only one node may be compromised (the *single* adversary) (Syverson et al., 2000). It must be noted that although in practice it often occurs for the attackers to become omnipresent adversaries in local area networks (the attacker manages to obtain access to a network administrator account), becoming an omnipresent adversary in large networks or in the Internet, is commonly considered as low probable or even infeasible due to the required resources and the technical difficulty in breaking into various differently protected systems. In the most optimistic scenarios, attackers would manage to observe *k* selected nodes (which they chose as the nodes of particular interest) in large networks.
- Active/passive – An *active* adversary can arbitrarily modify the computations and messages (by adding and deleting) whereas a *passive* adversary can only listen (Raymond, 2001).
- *Static/adaptive* – *Static* adversaries choose the resources they compromise before the protocol starts and are not able to change them once the protocol has started. *Adaptive* adversaries are allowed to change the resources they control while the protocol is being executed (Lindell, 2010; Raymond, 2001). They can, for example, 'follow' a message (Raymond, 2001).
- *Hybrids and alliances* – *Hybrids* and *alliances* of attackers may occur, such as external-active or colluding internal and external (see Figure 1c). Syverson et al. (Syverson et al., 2000), for example, distinguish between a *multiple* adversary and a *roving* adversary, which are subsequently: k-listening static and k-listening adaptive adversary.

For the matter of convenience, the k-listening, roving and omnipresent adversaries will be altogether called *k-present adversaries*, where, the *k* number is equal to the number of nodes in the network for omnipresent adversaries.

Figure 1. Types of adversaries performing traffic analysis: (a) external adversary, (b) internal adversary, and (c) hybrid (colluding internal and external) adversary.





(**c**)

ATTACKS

Time correlation

Time correlation, together with content correlation (see Section 3.2) are the primary traffic analysis attacks. Time correlation is the activity of linking incoming and outgoing agents based on time relations. In traditional networks, it is easy to trace messages that pass intermediate nodes, because the first-in-first-out rule is usually satisfied. Also for agents, if they only traverse an intermediate network node where they do not perform any tasks, this condition would apply. Time correlation may be performed by k-listening or omnipresent adversaries, or multiple colluding roving adversaries (spying agents) – k-present adversaries (see the previous section). Either external or internal.

This attack can be mitigated by modifying the time relationships between incoming and outgoing agents. This can be done by stopping agents at each node until a particular number of agents arrives at the node and forming a *batch* of agents. It must be assured that the messages in the batch are not located in the same order as they arrived at the node. This requirement is satisfied through *reordering* of agents – changing their location in the batch. This method was already utilised by Chaum in the first mix (Chaum, 1981). Another method relies on delaying agents for a random time at each network node, where the time of delay is assigned according to probabilistic calculations based on statistical characteristics of network traffic, to make sure that this will lead to an expected reordering of agents (Kesdogan et al., 1998).

Correlation based on distinguishing features of agents (including content correlation attacks)

The attack takes advantage of the fact that it is easy to recognise and trace agents through subsequent network locations, as long as the agents are easily distinguishable from other agents. For instance, such distinguishing features can be a non-uniform size (in terms of the number of packets) of an agent or a characteristic content of the data held by the agent. The attacks that target the latter features are also called *content correlation attacks* (Gülcü & Tsudik, 1996; Raymond, 2001). The attack may be performed by k-present adversaries, either external or internal. There is also an active version of the attack (see Section 3.11). In this attack, the attacker follows an agent, thus the more nodes the attacker can observe, the more likely they will discover the agent's base and destination.

To prevent this attack agents should be indistinguishable. To achieve this, firstly, the standard size of agents must be induced in an agent system (*agent size unification*). Afterwards, all the data that render an agent distinguishable from others should be obfuscated (*data obfuscation* or *content unification*).

This obfuscation may be realised via one-to-one hashing (permutations), encryption or any other method which assures that the obfuscated data are different before and after traversing each network node. An interesting method of agent data obfuscation was proposed by (Hohl, 1998). Unfortunately, size unification of all mobile agents in the target environment leads to either costly redundancies if the size is large, or inconvenient boundaries, if the size is too small, imposing the distribution of data between more than one agent. Thus at least groups of agents of the same size should be introduced to the environment.

Brute force attack

Suppose a deserted network area in which only a few agents roam. There, an attacker may learn about agents' bases and destinations just through tracing them, i.e. through observing an agent entering a network node and then following all outgoing agents. This approach is characterised as a *brute force attack* (Raymond, 2001). The attack may be very successful in such deserted environments, especially when agents follow separate paths and they do not meet at the same node at the same time. On the contrary, if a network is very crowded and agents frequently traverse the same network locations where they stay at the same time, it is difficult to distinguish agents as it is difficult to correlate agents visiting and leaving the network nodes. This observation led to the idea of the *mix* (Chaum, 1981). Roughly speaking, a mix is an intermediate node that needs to be passed by a message to lose its correlative relationships. A mix collects incoming messages into n-sized batches, detects and removes replays of the messages, decrypts each message to obtain the address of the next location and the message to be sent to it, reorders messages in the batch and outputs the batch (Chaum, 1981). Also in 'crowded' environments, an attacker may approach applying the brute force attack in hope that the situation in the network changes and that in certain network areas some separate paths could be observed. In this case, it is important to perform the attack continuously for a long time. The attacker performing brute force attack must be able to trace multiple agents if it occurs that the particular agent visits one network node and then many agents leave it. Thus the attack is attributed to external or internal k-present adversaries.

As far as protection from this attack is concerned, guaranteeing that network nodes are always visited by a certain number of agents and more than one leaves a node at the same moment, decreases the probability of linking a base and a destination exponentially with the length of the route. However, the network may be little-occupied (low traffic) and at a particular instant, the agent was the only one that traversed a given route. Then it is entirely exposed to an adversary. In traditional networks, to avoid this, dummy traffic is usually introduced i.e. redundant messages are sent from particular locations to another, to render the network always occupied. The more dummy messages are sent, the less probable is to correlate certain messages. This is a costly option, overloading the network, narrowing its bandwidth, introducing redundant computations. Concerning mobile agents, the situation improves, because all nodes are equal in the sense that there is no distinction between mixes and ordinary nodes (all nodes naturally become mixes). So the attacker cannot easily detect which node is the destination one. They might guess that it is the last one before returning to the base, but it is sufficient forcing the agent to traverse additional network nodes on its way back to prevent such inferences (redundant agent migration). Thus if there is only one destination for a particular agent, it is very difficult to deduce which one is it.

To prevent discovering an agent's base station, one option is to introduce redundant agents which start their lifecycle on random network nodes and then roam around the network waiting for a task. When a user needs an agent they pick it up from the non-occupied roaming agents rather than launch a new one. They assign a task to the agent and let the agent migrate to complete it. After the agent

comes back with the results, it is released to finish its lifecycle. Because the agent started its lifecycle earlier on a network node different to the one where it is being picked up, this previous node is the base node and not the one where the agent was employed. Note that these redundant agents differ from the 'dummy' ones since they finally are being used. Also, their number can be controlled and adjusted according to agents' usage statistics. This protective mechanism is not as strong as the introduction of dummy traffic, because the probability decreases linearly with the route length, but it is much less costly and may be sufficient, especially if the network is naturally loaded with normal agent traffic. At the same time it is worth noting that in practice, for an attacker, it is relatively difficult to perform the brute force attack besides the deserted network case. First of all, it is rather unlikely to obtain access to k-locations (or all of them) in the network or to introduce a high number of spying agents. Second, the observation needs to be performed for a long time. Finally, it is not facile to find the often not obvious correlations between the observations.

Replay attacks

Replay attacks are also referred to as *tagging through shadowing*, a specific instance of agent tagging attacks. They are described in Section 3.11.

Timing attacks

If traversing a route of a particular length always takes an agent a specific time, then this can be used by an adversary to correlate agents. For instance, (Raymond, 2001) assumes a scenario where passing a set of network locations takes one unit a second and a second unit – three seconds. Then if two arrivals are observed in the network at 0:00 and 0:01 and departures at 0:03 and 0:02, it is possible to discover which unit passed which route.

Regarding the types of adversaries that are able to perform the attack, two attack stages must be distinguished. In the first stage, the attacker measures the times needed for covering different routes. To obtain comprehensive information about the time, the attacker must compromise multiple network locations (to observe which route is being passed when the time is measured). Thus, this part of the attack is performed by external or internal k-present adversaries. Once the attacker obtains the necessary data, then it is sufficient to locate a particular network node which the attacker assumes to be a base node for a particular agent. Then, the adversary measures the time between the agent leaves and returns to the node, learning by this which route the agent passed, and possibly what was its destination. At this stage, it is sufficient for the adversary to be a single, passive adversary, internal or external. It is important to note, that the attacker is already in possession of significant information if they can call a network node the base for a particular agent.

Timing attacks against mobile agents are less successful due to the difficulty in identifying the destination (see 3.3). An attacker performing this attack receives the information about the route the agent passed so only an indication of several potential destinations. To increase this already high level of anonymity it is worth considering the introduction of random delays of agents' visits to network nodes or batch processing (see Sections 3.1 and 3.6). Note, that in practical agent environments, in which agents roam and perform tasks, random delays may naturally result from the tasks. Furthermore, if the redundant migration or redundant agents (see Section 3.3) are applied, then an attacker is not able to learn neither about the base nor the destination.

Flooding attack (a.k.a. spam attack, node flushing attack, n-1 attack, isolate & identify attack)

As it was already mentioned in Section 3.3, an effective preventive action is to cause multiple agents to visit the same network node simultaneously, so it becomes difficult to correlate the incoming agents and the outgoing ones. A method that aims at imposing this feature in traditional message communication is the batch processing (see Section 3.1). The method relies on deliberately collecting *n* messages before they are released (*flushed* – see (Raymond, 2001)) from a network node. This results in messages leaving the nodes in *n*-sized batches. Reordering of messages is key in this method, as it distorts the time-relations between incoming and outgoing messages. The method hinders performing brute force attacks, as well as timing and time correlation attacks. The drawback is message delays and uneven network traffic ('waves' of messages).

An attacker response to this protection is the *n-1 attack* (other popular names are: *flooding attack*, *node flushing attack* and *spam attack*). An adversary 'fills' the mix with *n-1* of their messages, allowing only one foreign message to join the batch. Then, when the batch of *n* messages leaves the mix, it is trivial for the adversary to separate the message which they observed. This attack is primarily performed by external adversaries. As in previous cases, to trace an agent the adversary must observe multiple nodes, thus they must be k-present.

To prevent the n-1 attack, it must be assured that either the adversary is not able to recognise their *n-1* agents after they leave the network node or to prevent the attacker from delivering the agents to the node. Raymond (Raymond, 2001) proposes encrypting the traffic between network nodes, which according to the author should result in that an attacker loses the ability to easily recognise their flooding agents. However, this is only partially true. If the attacker assures that all their flooding agents go to a particular network node after leaving the observed one, then they are distinguishable unless the observed agent also goes to the node (which does not interfere with the main aim of the adversary, which is to recognise where the agent goes). At the same time, the encryption disallows the adversary to recognise their agents through examination of their data. In place of the encryption proposed by Raymond, also other data obfuscation methods discussed in Section 3.2 can be applied. Another solution (used in Stop-and-Go mixes (Kesdogan et al., 1998)) is to force agents to wait at the network node for a random amount of time. Then, even if the adversary had flooded the node with their *n-1* agents, they would not leave the node in the same batch. Instead, *k* agents that do not belong to the adversary will be in the batch, which hinders correlating incoming and outgoing agents. This countermeasure appears to be effective, for the cost of delays in agent communication. To summarise, to protect mobile agents from the flooding attacks, the agent data obfuscation described in Section 3.2 together with agent delaying, should be used instead of n-batches.

Contextual attacks

Contextual attacks refer to the communication habits of users. They are performed by k-present adversaries. There are three types of attacks in this group (Raymond, 2001): communication pattern attacks, packet counting attacks and intersection attacks. *Communication pattern attacks* aim at observing users' habits in using network services. For instance, suppose a company employee who prefers to work late evenings. If the person is the only one using the company network at a particular time then it is not difficult to connect agents active in the network with the person. As far as *packet counting attacks* are concerned, they are targeting the situation when a user launches an agent of a characteristic, distinctive size (in terms of the number of packets). This can occur, for instance, when a file is attached to the agent. When performing an *intersection attack*, an adversary observes network traffic and stepwise narrows the range of possible interlocutors (as described in (Berthold et al., 2000)). Imagine an agent travelling twice from a base A to a destination B, each time passing through

completely different network nodes. If an adversary observes these two trips, they notice that the only network locations in common are A and B, which makes them good candidates for interlocutors. Unfortunately, this attack undermines the protection based on using different routes each time an agent goes to the same destination that proves effective for untraceability denial of service attacks (see Section 3.8).

A method for avoiding packet counting attacks, called *size unification* was already described in Section 3.2. In general, contextual attacks are difficult to protect from, because they rely on factors that are beyond the control of system designers and administrators, namely the unpredictable users' behaviours. The only way to prevent the attacks is not performing characteristic activities, nor exhibiting distinguishing habits. Appropriate security and privacy training and awareness raising play an important role here to make users conscious of the fact that by performing characteristic activities, they become vulnerable to tracing.

Untraceability denial of service attacks

The following description focuses on the *denial of service* (DoS) attacks that aim at compromising *untraceability*. An attacker disrupts some intermediary nodes, counting on the fact that it would force a user who usually communicates via them, to change his or her behaviour. Though this attack is effective in conventional message-based networks, in the case of mobile agents it appears to be unsuccessful, as agents arbitrarily choose a different route each time they roam. When they encounter a non-functioning network node, they simply omit it and choose another. This feature is further described in Section 3.11. The attack can be performed by active (static or adaptive, external or internal) and k-present adversaries.

As a countermeasure to this attack, it must be assured that an agent does not behave abnormally when encountering a compromised network node. It means that the agent should pass the entire route as in the normal situation (instead of promptly returning to its origin to report the damage). However, it is not required from the agent to continue the realisation of its task. The agent may stop its goal-oriented activities. The objective is to prevent an outside observer to notice any difference in the agent's behaviour. To achieve this, a *failure-neutral behaviour* is proposed. The failure-neutral behaviour is an agent behaviour that in the face of a system failure remains indistinguishable for an external observer from the behaviour in the normal system operation. Failure-neutral behaviour is a less demanding feature than fault tolerance, however, it can be perceived as a part of it. Fault-tolerance refers to the more restricted property of systems, requiring them to continue their proper operation if the failure occurred. For failure-neutral behaviour it is not required from an entity to remain properly operating, internal functionalities may be disrupted.

Active attacks exploiting user reactions

The attacks aim at provoking behaviours of agents or agent owners that would facilitate their tracing. For instance, an agent can be intercepted, cloned and the clones sent to all possible recipients, assuming that the original recipient would behave differently from others. As in the case of untraceability denial of service attacks, these attacks are performed by active (static or adaptive, external or internal) and k-present adversaries.

One solution to prevent this type of attacks is to assure that agents go further than their base network nodes when returning (*redundant agent migration*). This idea was already described in Section 3.3. Also completing the whole route even in unusual situations (in this case – that the intended destination

appears not to be the real one) – the failure-neutral behaviour – as in the case of DoS attacks (see Section 3.8), can be applied. It is very difficult to imagine all attacks aiming at provoking characteristic user reactions, and thus it is very difficult, if not impossible, to propose one effective protection method. The only realistic approach is based on learning from experience – known from the area of malware protection, where countermeasures are developed immediately after a new attack is discovered. The time between the threat discovery and the release of a remedy plays a crucial role here. Also, security and privacy awareness and knowledge exchange are important measures that help in protecting from these attacks.

Agent delaying

When performing the *agent delaying attack*, an attacker stops an agent for a specific purpose or until a certain condition is satisfied. For instance, an agent may be delayed to see if potential destinations will be visited by other agents, or until sufficient network resources are obtained by an adversary, or the network becomes easier to monitor for the adversary, etc. The attack is performed by active (static or adaptive, external or internal) adversaries to facilitate further investigations of k-present adversaries.

To protect from this attack, administrators should consider introducing *authenticated timing information* – timestamps of the arrival and the departure at network nodes securely encapsulated into agents. For instance time windows, as in Stop-and-go-MIX'es (Kesdogan et al., 1998) can be used.

Agent tagging

In *agent tagging*, an adversary purportedly 'marks' an agent (alters the agent's data or behaviour to make the agent distinctive) to facilitate its tracing. There are three types of attacks identified in this group: *tagging data*, *tagging through delaying* and *tagging through shadowing*. The first and the most intuitive type of tagging attacks relies on changing the agent's data, so the agent was easily distinguished from other agents. The attacker may add some characteristic data to the agent but also remove or change existing data. The feasibility of tagging-data attacks is very dependent on a particular network and information system and should be discussed concerning the real environment. In most cases, network protocol characteristics prevent such tagging on the network layer, which makes the attack unavailable to an external attacker. Thus, in the first tagging stage the attack must be performed by active internal adversaries. Once the agent is tagged, its tracing, as in the case of correlation-based on distinguishing features (see Section 3.2) belongs to roving adversaries, k-listening or omnipresent adversaries, either external or internal.

Tagging through delaying attack can be perceived as a specific version of the agent delaying attack (see Section 3.10) that aims at making an agent distinguishable in the network. During the attack, an attacker forces an agent to stop at each network node for a specific, characteristic time. After this, it is possible to correlate arrivals and departures of such a tagged agent to and from a network location. The attack is available to active, internal/external attackers, with the ability to observe the agent at multiple nodes (thus k-present). The *tagging through shadowing* attack (more often known as *replay attack* (Gülcü & Tsudik, 1996)) is based on intercepting an agent and copying it. After this, *k* copies of the agent traverse the same route. This attack is effective in traditional networks and in the mobile agent networks where agents plan their routes using a deterministic algorithm. In these systems, the copies of messages/agents will follow the same route as the original. In the agent networks where agents have freedom in autonomously choosing their route, the attack appears to be ineffective because each of the copies autonomously chooses its route. Thus, this is important to assure that

agents select their routes in a non-deterministic way (*non-deterministic routing*). Deterministic routing forms another serious vulnerability – if adversaries can discover its algorithm they could predict agent routes.

Tagging through delaying attacks can be avoided through early detection of alterations of the agent's code (*tamper detection*), or optimally, by not allowing an adversary to change the code of the agent (*tamper-resistance*). The methods aiming at detecting agent data alterations include *agent tampering detection by storage jamming* (Meadows, 1997) and various schemas based on hashing agents' data. In the latter case, it is important to assure that the hashes are verified at each network node, so the alteration was detected before the agent covered its route. Preventing agent data from being altered is very difficult if not infeasible. So far no method has been proposed that would effectively resolve this problem. Protecting from tagging data attacks is similar to protecting from tagging through delaying attacks. For both of them, it is crucial either to detect or not to allow the alteration. In the case of tagging data attacks the protection task is slightly easier since only the data carried by an agent – a more static part of an agent (comparing to code) – are processed. The basic technique for preventing tagging through shadowing attacks is replaying' detection which uses one of the following techniques: sequence numbers, random numbers or data and timestamps (Gülcü & Tsudik, 1996) or keeping the record of previous agents (Chaum, 1981). On the other hand, if agents choose their routes in a non-deterministic way, the copies of the agents do not follow the same way as the originals, which makes the attack ineffective.

Corrupted-party attacks (a.k.a. 'sting' attacks)

Corrupted-party attacks rely on taking over either the agent's base or the destination, followed by masquerading as a genuine party of communication. For instance, an attacker could set up a home site with illicit-looking content and observe agents' visits to the site (Raymond, 2001)). Another scenario includes an attacker querying sites, and observe responses, behaving as an initiator of the communication. The observations of queries and responses can be performed because the attack assumes that an adversary can encode some indicative information into an agent, and then follow the tagged agent as in the case of agent tagging attacks (see Section 3.11). Thus, the corrupted party attacks are primarily attributed to internal active k-present adversaries. The attacks prove to be adequate also for active internal adversaries who are not able to observe larger areas of the network (in practice – to single adversaries). In their case, the crucial role plays the ability to involve a user located on the other side of communication, into the conversation, so an agent after returning to the user would come again to the adversary. Then the user needs to be induced to release some identifying information about them or to make the agent obtaining address data of the base network node.

The most effective protection method is to prevent tagging an agent at the destination (tamper resistance, see Section 3.11). However, this is only possible in situations when agents do not collect any data at the destination. If an agent gets involved in a conversation, then it is difficult to detect if the exchanged data may serve for the tagging purpose. One approach could be to detect the data which are not used at the user's side but returned to the destination (tamper detection, see Section 3.11). In the case of single adversaries aiming at forcing a user or an agent to release identifying information, it is possible to mitigate the attack at the technical level by assuring network nodes not to disclose their identifying data to agents. However, as it was already described for contextual attacks (see Section 3.7) – it is impossible to control users' behaviours. Thus, as in the case of contextual attacks, the users need to take appropriate actions to protect themselves. Here, security and privacy awareness plays a crucial role.

CONCLUSIONS

Nowadays, when mobile agents are applied to newly emergent domains such as Big Data, the Internet of Things, Smart Grids and others, appropriate risks must be considered when developing the applications. The analysis presented in this paper shows that an adversary willing to violate the privacy of agents' owners, has quite an impressive portfolio of attacks at his or her disposal. This includes various types of correlations, observing the behavioural patterns of agents' owners or to involving them into interactions with attackers. Besides, the attackers can assume various configurations, including, for instance, the surrounding of a particular network node, or compromising a host to take it over. Fortunately, several countermeasures have been proposed. Also, not all of the attacks are practically feasible. For instance, the attack scenarios devoted to k-present adversaries are difficult to implement in large-scale environments consisting of hundreds of network nodes because they require practically unreachable resources from an attacker. These considerations need to be incorporated into appropriate threat models during cybersecurity management and when designing and developing new mobile agent-based applications.

Another important question concerns the ethical component associated with the application of anonymisation mechanisms. As much as the technology is devoted to protecting the fundamental right of privacy, it can be also used for malicious purposes. Proxies, onion routing or encrypted communications (Koch, 2019; Montieri et al., 2020) are primary instruments that enable immersing in the anonymous world of the Dark Web and engage in deleterious or illegal activities including criminal or terroristic. Essential questions arise that regard the extent to which the technologies can be applied. One of the challenges is that the technology may lower states' ability to access the necessary information when it is justified to protect citizens. This is because it hinders collecting intelligence used to detect, locate and prevent a range of potential threats. Yet, it is a state's ethical obligation to protect its citizens. On the other hand, the processed data may represent the most intimate and private values to the individual (Bellaby, 2018). Thus, there are important concerns related to the individual's right to establish the technological barriers, the necessary circumstances, and, how the state should respond. Besides, it needs to be decided which forms of state intervention are justified. Whether the technology should be completely prohibited or more gradual measures need to be introduced.

KEYWORDS: mobile agents, anonymity, privacy, ethics, traffic analysis, tracking, autonomous systems, Internet of Things (IoT), Big Data, Dark Web, privacy management, cybersecurity management, organisation management.

REFERENCES

Bellaby, R. W. (2018). Going dark: anonymising technology in cyberspace. *Ethics and Information Technology*, *20*(3), 189-204. https://doi.org/10.1007/s10676-018-9458-4

Berthold, O., Pfitzmann, A., & Standtke, R. (2000). The disadvantages of free MIX routes and how to overcome them. In H. Federrath (Ed.), *Proceedings of Designing Privacy Enhancing Technologies: Workshop on Design Issues in Anonymity and Unobservability* (Vol. 2009, pp. 30-45). Springer-Verlag New York, Inc.

Blaze, M., Ioannidis, J., Keromytis, A. D., Malkin, T., & Rubin, A. (2005). *WAR: Wireless Anonymous Routing* (pp. 218-232). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11542322_27

Bouchemal, N., & Maamri, R. (2016). *CAPMA: Clone agent to protect mobile agents in dynamic environments*. 17-21. https://doi.org/10.1109/ICAASE.2016.7843865

Calvaresi, D., Calbimonte, J.-P., Debusson, F., Najjar, A., & Schumacher, M. (2019). Social Network Chatbots for Smoking Cessation: Agent and Multi-Agent Frameworks. *IEEE/WIC/ACM International Conference on Web Intelligence*, 286-292. https://doi.org/10.1145/3350546.3352532

Chaum, D. (1981). Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, *4*(2).

Choi, H., Park, J., & Jung, Y. (2018). The role of privacy fatigue in online privacy behavior. *Computers in Human Behavior*, *81*, 42-51. https://doi.org/10.1016/j.chb.2017.12.001

Cottrell, L. (1995). *Mixmaster and Remailer Attacks*.

Dasgupta, D., & Brian, H. (2001). Mobile security agents for network traffic analysis. *Proceedings DARPA Information Survivability Conference and Exposition II. DISCEX'01*, *2*, 332-340. https://doi.org/10.1109/DISCEX.2001.932184

Dolev, S., & Ostrobsky, R. (2000). Xor-trees for efficient anonymous multicast and reception. *ACM Transactions on Information Systems Secururity*, *3*(2), 63-84.

Goldberg, I., & Wagner, D. (1998). TAZ servers and the rewebber network: Enabling anonymous publishing on the world wide web. *First Monday*, *3*(4).

Gray, R. S., Cybenko, G., Kotz, D., Peterson, R. A., & Rus, D. (2002). D'Agents: Applications and performance of a mobile-agent system. *Software, Practice and Experience*, *32*(6), 543-573.

Gray, R. S., Kotz, D., Cybenko, G., & Rus, D. (2000). *Mobile Agents: Motivations and State-of-the-Art Systems* (Issues TR2000-365).

Gülcü, C., & Tsudik, G. (1996). Mixing Email with Babel. *Proceedings of the 1996 Symposium on Network and Distributed System Security (SNDSS '96)*, 2. citeseer.nj.nec.com/2254.html

Hohl, F. (1998). Time Limited Blackbox Security: Protecting Mobile Agents From Malicious Hosts. In G. Vigna (Ed.), *Lecture Notes in Computer Science* (Vol. 1419, pp. 92-113). Springer-Verlag New York, Inc.

Isern, D., & Moreno, A. (2016). A Systematic Literature Review of Agents Applied in Healthcare. *Journal of Medical Systems*, *40*(2), 43. https://doi.org/10.1007/s10916-015-0376-2

Jolly, P. K., & Batra, S. (2019). Security against Attacks and Malicious Code Execution in Mobile Agent Using IBF-CPABE Protocol. *Wirel. Pers. Commun.*, *107*(2), 1155-1169. https://doi.org/10.1007/s11277-019-06329-7

Kampik, T., Malhi, A., & Framling, K. (2019). Agent-Based Business Process Orchestration for IoT. *IEEE/WIC/ACM International Conference on Web Intelligence*, 393-397. https://doi.org/10.1145/3350546.3352554

Kem, O., & Ksontini, F. (2019). A Multi-Agent System for Energy Management in a Dynamic and Open Environment: Architecture and Optimisation. *IEEE/WIC/ACM International Conference on Web Intelligence*, 348-352. https://doi.org/10.1145/3350546.3352545

Kesdogan, D., Egner, J., & Büschkes, R. (1998). Stop-and-Go MIXes: Providing Probabilistic Anonymity in an Open System. *Proceedings of Information Hiding Workshop (IH 1998)*, *1525*.

Koch, R. (2019). Hidden in the Shadow: The Dark Web - A Growing Risk for Military Operations? In G. Minarik, T and Alatalu, S and Biondi, S and Signoretti, M and Tolga, I and Visky (Ed.), *1th International Conference On Cyber Conflict (CYCON): Silent Battle* (pp. 267-290). IEEE.

Kulesza, K., Kotulski, Z., & Kulesza, K. (n.d.). *On Mobile Agents Anonymity; formulating traffic analysis problem*.

Kulesza, K., Kotulski, Z., & Kulesza, K. (2006). On Mobile Agents Resistance to Traffic Analysis. *Electronic Notes in Theoretical Computer Science*, *142*, 181-193. https://doi.org/10.1016/J.ENTCS.2004.12.044

Lindell, Y. (2010). *Foundations of Cryptography 89-856*. http://u.cs.biu.ac.il/~lindell/89-856/complete-89-856.pdf

Lopez, J., Rios, R., Bao, F., & Wang, G. (2017). Evolving privacy: From sensors to the Internet of Things. *Future Generation Computer Systems*, *75*, 46-57. https://doi.org/10.1016/J.FUTURE.2017.04.045

Luck, M., McBurney, P., & Preist, C. (2003). *Agent Technology: Enabling Next Generation Computing (A Roadmap for Agent Based Computing)*. AgentLink.

Madkour, M. A., Eassa, F., Ali, A., & Qayyum, N. (2014). Securing mobile-agent-based systems against malicious hosts. *World Applied Sciences Journal*, *29*, 287-297. https://doi.org/10.5829/idosi.wasj.2014.29.02.1561

Mayowa, O., M., F., Olabiyisi, S., Omidiora, E., & Fawole, A. (2016). A Survey on Migration Process of Mobile Agent. *Proceedings of the World Congress on Engineering and Computer Science*.

Meadows, C. (1997). Detecting Attacks on Mobile Agents. *Proceedings of Foundations for Secure Mobile Code Workshop*, 64-65. citeseer.ist.psu.edu/meadows97detecting.html

Montieri, A., Ciuonzo, D., Aceto, G., & Pescapé, A. (2020). Anonymity Services Tor, I2P, JonDonym: Classifying in the Dark (Web). *IEEE Transactions on Dependable and Secure Computing*, *17*(3), 662-675. https://doi.org/10.1109/TDSC.2018.2804394

Pellungrini, R., Pappalardo, L., Pratesi, F., & Monreale, A. (2017). A Data Mining Approach to Assess Privacy Risk in Human Mobility Data. *ACM Transactions on Intelligent Systems and Technology*, *9*(3), 1-27. https://doi.org/10.1145/3106774

Raymond, J.-F. (2001). Traffic Analysis: Protocols, Attacks, Design Issues and Open Problems. In H. Federrath (Ed.), *Designing Privacy Enhancing Technologies: Proceedings of International Workshop on Design Issues in Anonymity and Unobservability* (pp. 10-29). Springer-Verlag New York, Inc.

Reiter, M., & Rubin, A. (1998). Crowds: Anonymity for Web Transactions. *ACM Transactions on Information and System Security*, *1*(1). https://doi.org/10.1145/290163.290168

Sanae, H., Laassiri, J., & Berguig, Y. (2019). *Security Requirements and Model for Mobile Agent Authentication* (pp. 179-189). https://doi.org/10.1007/978-981-13-8614-5_11

Sobh, T., & Elleithy, K. (Eds.). (2015). *Innovations and Advances in Computing, Informatics, Systems Sciences, Networking and Engineering* (Vol. 313). Springer International Publishing. https://doi.org/10.1007/978-3-319-06773-5

Syverson, P., Tsudik, G., Reed, M., & Landwehr, C. (2000). Towards an analysis of onion routing security. In H. Federrath (Ed.), *Proceedings of Designing Privacy Enhancing Technologies: Workshop on Design Issues in Anonymity and Unobservability* (pp. 96–114). Springer-Verlag New York, Inc.

Urra, O., Ilarri, S., & Trillo-Lado, R. (2017). An approach driven by mobile agents for data management in vehicular networks. *Information Sciences*, *381*, 55-77. https://doi.org/10.1016/J.INS.2016.11.007

Xia, H., Wang, Y., Huang, Y., & Shah, A. (2017). "Our Privacy Needs to be Protected at All Costs." *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1-22. https://doi.org/10.1145/3134748

Yang, Y., Lu, W., Xing, W., Wang, L., Che, X., & Chen, L. (2017). Detecting and resolving deadlocks in mobile agent systems. *Journal of Visual Languages & Computing*, *42*, 23-30. https://doi.org/10.1016/J.JVLC.2017.08.002

Zrari, C., Hachicha, H., & Ghedira, K. (2015). Agent's Security During Communication in Mobile Agents System. *Procedia Computer Science*, *60*, 17-26. https://doi.org/10.1016/J.PROCS.2015.08.100

AN EMPATHIC LEARNING PEDAGOGY MODEL FOR AN EXPERIENTIAL PROJECT MANAGEMENT

Shalini Kesar, James Pollard

Southern Utah University (USA)

kesar@suu.edu; jamespollard1@suu.edu

INTRODUCTION

The motivation for this research started with the belief that core skills such as teamwork, communication, professionalism and ethics are part of experiential learning pedagogy. These skills in turn prepare the students to deal with challenges faced in today's technology related businesses. The author (instructor) participated as an attendee as well as a speaker in many workshops and seminars associated with teaching and cultivating empathy during the COVID-19 challenging environment. For example, "Why Empathetic Teaching Matters: Going the Extra Mile" where the goal was to discuss empathy as a powerful tool that can help you better understand what's driving the students' behaviour. It also discussed different strategies to help connect and work through difficult moments together as an educator and student (Paradigm Education Solutions, 2021). The strategies on how to cultivate an online learning community for students provided examples that the authors believe led to redesigning the teaching style that led to "going the extra mile to create social connections that will help students stay connected, motivated, and engaged in their learning" (Paradigm Educations Solution, 2021).

This paper reflects on experiential learning pedagogy for a project management class that was modified during the beginning of the COVID-19 outbreak. Further, it discusses how the instructor used the eight principles of National Society for Experiential Education (NSEE) framework and modified them to cultivate an empathic learning classroom environment. It begins by an explanation of NSEE framework, followed by the impact of COVID-19. It outlines the project and how the NSEE framework's principles were modified or/and revisited to cultivate an empathic learning pedagogy. Finally, it discusses some concluding remarks including lessons learned. The contribution of this paper is that these suggestions can be used by other instructors while designing experiential classrooms. This is important and has significantly impacted on how educators as well students view what constitutes an experiential classroom environment in the new "normal" during and post pandemic times.

NSEE FRAMEOWRK

The National Society for Experiential Education (NSEE) framework the varied roles and responsibilities represented in the field of experiential education. Founded in 1971. The NSEE model continued to provide an experiential leaning environment for the students. Their mission is to cultivate educators who effectively use experiential education as an integral part of personal, professional, civic and global learning. It advocates the use of experiential learning throughout the educational system; to disseminate principles of best practices and innovations in the field; to encourage the development of research and theory related to experiential learning; to support the growth and leadership of experiential educators; and to create organizational partnerships with the community. It is considered as an "open and pluralistic society of individuals and institutions dedicated to mutual learning and support across the varied roles and responsibilities represented in the field of experiential education"

(NSEE, 2011). The Board of Directors, staff, and membership of NSEE have been governed by policies and practices that guide ethical actions, relationships, and decisions. This research was also guided by Principles of Best Practices as well as ethical principles outlined for NSEE. These principles are guided by the Statement of Shared Ethical Principles (Council for the Advancement of Standards in Higher Education), National Education Association, American Association of University Professors, and Code of Ethics for Education Abroad (NSEE, 2011).

The National Society for Experiential Education's framework proved to be useful in developing the curriculum. Eight weeks into the semester, the global pandemic resulted in the cancellation of face to face course structure which was being replaced with remote classes. Subsequently, deliverables of the projects, assignments, weekly reporting, and presentations were modified to adjust to remote teaching practices.

The NSEE model continued to provide an experiential learning environment for the students. The mission of the NSEE is to cultivate educators who effectively use experiential education as an integral part of personal, professional, civic and global learning. The National Society for Experiential Education society advocates the use of experiential learning throughout the educational system; to disseminate principles of best practices and innovations in the field; to encourage the development of research and theory related to experiential learning; to support the growth and leadership of experiential educators; and to create organizational partnerships with the community. The eight Principles of Good Practice for All Experiential Learning Activities include: Intention; Preparedness and Planning; Authenticity; Reflection; Orientation and Training; Monitoring and Continuous Improvement; Assessment and Evaluation; and Acknowledgment. Given that this is an ongoing research, in this paper, this NSEE framework is used as a methodology to reflect on the approach used to cultivate an empathic learning experience.

Intention: The first principle focuses on the creating a pedagogy that can demonstrate the purposefulness to enable experience to become knowledge. Consequently, it is more than just outlining the goals and objectives, and activities that define the experience.

Preparedness and Planning: Based on the first principle above, the second step ensures participants enter the experience with sufficient foundation to support a successful experience. In this step, it is important that intentions are identified with the goals and objectives.

Authenticity: In this principle, the experience designed must have a real-world context and /or be useful and meaningful in reference to an applied setting or situation. This is unique within this principle.

Reflection: This step/principle helps to transform simple experience to a learning experience. These activities designed can create knowledge that the learner can internalized. Furthermore, it allows the instructor to reflect on the assumptions and hypotheses about the outcomes of decisions and actions taken, along the outcomes against past learning and future implications. This reflective process is integral to all phases of experiential learning, from identifying intention and choosing the experience, to considering preconceptions and observing how they change as the experience unfolds.

Orientation and Training: This principle adds value of the experience to be accessible to both the learner and the learning facilitator(s), and other parties who are part of the experiential learning.

Monitoring and Continuous Improvement: This ensures that the experience, as it is in process, continues to provide the richest learning possible, while allowing responsibility and

accountability. The feedback will contribute towards the continuous improvement for an experiential learning experience.

Assessment and Evaluation: In this principle, the processes should be systematically documented with regard to initial intentions and quality outcomes. Assessment is a method to not only develop but to refine the initial goals and quality objectives identified.

Acknowledgment: The last principal is a recognition process of learning and its impact on all the parties involved throughout the experience. This principle is a form of a celebration of learning and helps provide closure and sustainability to the experience.

IMPACT OF COVID-19

The COVID-19, global pandemic has brought many challenges to the educational system to the world. Many reports and studies were conducted on the impact, coping strategies, and next steps on how to cope with the complicated COVID-19 situation. For example, forces, energy plants, sports facilities, and other civic institutions were shut down, which had a major impact in the local and regional economies. In addition, many employees in hospitals were placed at the front line of the local healthcare system. Education system was severely impacted particularly universities because they serve such a wide variety of functions. Educational institutions, where undergraduate and postgraduate students learning platform and faculty teaching and conducting research changed to a complete remote classroom environment (Illanes et al., 2020).

Specifically, in the education context, many issues of concerns surfaced as the global pandemic led to a new “normal” situation. The new “normal” led to all classes in schools, colleges and universities to be conducted remotely from March 2020. As the education system transited into to a remote environment, students started to question, for example, “How will academic credit be determined?; “Will I get reimbursed for unused room and board?; “ Will there be a commencement ceremony?”; “How will this affect my athletic scholarship?”; and “ Can I even stay in the country if my student visa is revoked?” (Illanes et al., 2020). In addition, accessibility and institution resources also making part of the discussion topics. One report stated, “Students from privileged backgrounds, supported by their parents and eager and able to learn, could not find their way past closed school doors to alternative learning opportunities. Those from disadvantaged backgrounds often remained shut out when their schools shut down. This crisis has exposed the many inadequacies and inequities in our education systems – from access to the broadband and computers needed for online education, and the supportive environments needed to focus on learning, up to the misalignment between resources and needs.” (Schleicher, 2020). By the end of March, all 46 countries in Europe had closed some or all of their schools (Schleicher, 2020). This and other similar reports clearly indicate that pedagogies need to reflect on specific needs of learning environment and allow pedagogy to adapt to the changing environment associated with the developing effects of the pandemic.

PROJECT MANAGEMENT

The Information Systems (IS) project management class gives senior-year students an opportunity to manage a major information systems development/enhancement project, in which they apply what they have learned in various other courses to a single project. The emphasis is on enterprise-level project management. The instructor encouraged guest lectures attend the course from different departments, including English, communication, and law. This provided background information and learning skills about technical writing, communication and presentation skill and related the

importance of these skills in the context and working environment in which they will be exposed to after graduation. The class duration was fifteen weeks and was conducted in the last semester before the students graduated.

The class curriculum was designed using the NSEE's eight principles (as outlined above). Having said that, NSEE'S framework was modified to cultivate empathic learning remote classroom experience environment for all the class projects. The instructor revisited the eight principles after eight weeks of class to incorporate some changes to aspects of the principles with the intent to continue to provide an experiential learning experience for the students during the global pandemic challenging times.

One of the four main projects providing focus for the students was the Design and Analysis Toolkit for Inventory and Monitoring (DATIM) which is an application being developed by the U.S. Forest Service. The United States Forest Service (USFS) is an agency of the United Sates Department of Agriculture that administers the nation's 154 national forests and 20 national grasslands, which encompass 193 million acres. Major divisions of the agency include the National Forest System, State and Private Forestry, Business Operations, and the Research and Development branch. The DATIM project is a collaborative effort between the National Forest System (NFS) and USFS Research & Development (R&D), Forest Inventory and Analysis (FIA), and Ecosystem Management Coordination (EMC) staff. The DATIM core team is comprised of both R&D and NFS staff from resource inventory and forest planning programs. The co-author, a Research Fellow at the University is the main collaborator with the development team of DATIM. He has been part of the DATIM cap-stone project for the last three years. The DATIM project has four modules: 1) Design Tool for Inventory and Monitoring (DTIM), which is designed to assist users in deter-mining objectives, questions, and metrics for monitoring plans; 2) Analysis Tool for Inventory and Monitoring (ATIM), which enables users to analyse vegetation data to derive estimates of current conditions and trends in the Forest and surrounding landscapes; 3) Spatial Intersection Tool (SIT), which enables users to add spatial attributes to DATIM datasets for use in ATIM; 4) DATIM Compilation System (DCS) enables users to add supplemental Forest Vegetation Simulator (FVS) attributes to DATIM datasets for use in ATIM.

CULTIVATING EMPATHIC PROJECT MANAGEMENT CLASSROOM ENVIROMENT

Eight weeks into the semester, the global pandemic resulted in the cancellation of face to face course structure which and was replaced with remotely conducted classes. Subsequently, deliverables of the projects, assignments, weekly reporting, and presentations were modified to adjust remote teaching.

Various studies highlighted the importance of extensive pedagogical and technological resources as faculty began to prepare an empathic approach with resilient teaching components for academic year 2020-2021. For example, Ravitch (2020) outlines five dimensions of flux pedagogy, where she describes integration of existing theories and pedagogical frameworks into the teaching components during this pandemic. In addition, studies indicated that boosting social connectedness and increasing helping behaviours, empathizing with others also improves the ability to regulate your emotions during times of stress. In general, feeling "empathy" allows to better manage the anxiety you are experiencing without feeling overwhelmed. In other words, empathy is an individual's capacity to understand the behaviour of others, to experience their feelings, and to express that understanding to them (for example, see Barrett-Lennard (1981) proposed a theory that empathy must involve "resonating" with another person's emotions (in Kagan & Schneider, 1987, p. 459), where a person "physiologically experiences the other person's affects" (Holm, 1996, p. 241). More recently, in his book, Loreman (2011), discussed empathy as a necessary ingredient of an effective pedagogy.

Given that NSEE's framework was used to design this class project, an empathic approach was incorporated in the teaching and experiential learning platform in current pedagogy. This helped students traverse complex systems during chaotic times, build relational trust, and welcome pedagogical flexibility as part of an experiential learning model. Beyond the synchronized electronic connection (via Zoom), the instructor believed it was important to connect with students with empathy, especially in times of anxiety and uncertainty. Although the NSEE framework was used, many other educational models were studied. For example, Sarkadi and Casmana (2020) state, "Learning empathy is one of the educational models to be able to educate students' characters, especially the character of empathy. This character is very important to be applied to students, because it can improve good relations between students. They can help each other, especially if students find it difficult in the distance learning process that is carried out at home. Students can help their friends if there are those who find it difficult, such as those who have difficulty connecting to the internet while the learning process is ongoing" (pg. 1043). The authors cultivated empathic learning pedagogy in the NSEE's eight principle by discussing the original design of the class followed by modification of NSEE framework in the project management course.

Intention: The first principle, and perhaps the most important point regarding the main intention, is the choice of capstone project. Based on the student learning goals of the capstone class, instructor and client (co-author) outlined a real business setting of the development of DATIM linked with testing the application for accessibility to people with disabilities. The Forest Service project was designed specifically to improve student learning with technical and core skills to meet the overall goal of testing the DATIM application. The students' task was to test manually the links (issues) in the DATIM application that could not be automated and were not in compliance with Section 508 guidelines for web applications for disabled persons. Section 508 of the Rehabilitation Act (29 U.S.C. § 794d), as amended by the Workforce Investment Act of 1998 (P.L. 105-220) requires federal agencies to develop, procure, maintain and use information and communications technology (ICT) that is accessible to people with disabilities - regardless of whether or not they work for the federal government. The US Access Board established the Section 508 standards that implement the law and provides the requirements for accessibility (https://www.epa.gov/accessibility/what-section-508). The student's project was designed to improve the accessibility testing process for the DATIM application

Preparedness and Planning: The NSEE practices indicate that it is valuable to ensure that participants (in this case the students) must ensure that they enter the experience with sufficient foundation to support a successful experience. As part of planning the class, the author (instructor) emailed students in November to provide an opportunity to participate in the early stages of the planning of the project. A few students showed interest and participated in a few meetings, the client provided some reading linked with the project. In the beginning of the semester, the client and team decided to visited the class to provide an overview of the project. As a result, six students decided to work on the DATIM project. Impact of COVID-19, led the instructor to rethinking the original intent and planning. This was crucial as students were facing many challenges. Some students had lost jobs, internships or job offers. This was true for other universities too. For example, Aucejo et al (2020) conducted a survey to reflect on the impact of COVID-19 on approximately 1,500 students from one of the largest public institutions in Arizona (United States). Result showed large negative effects across many dimensions. Due to COVID-19: 13% of students have delayed graduation, 40% lost a job, internship, or a job offer, and 29% expect to earn less at age 35. Keep in mind how change of modality in the middle of semester had negative effects across many dimensions. The instructor modified the class structure to include a thirty- minutes open discussion of challenges as they were faced.

Authenticity: Student project was linked with DATIM and testing Section 508 issues. This project experience was designed in an experiential learning, real-world context. Students continued working on their projects remotely. Like DATIM application website, all federal documents publicly available on government websites need to conform with Section 508 accessibility requirements.

One of the impacts of COVID-19 was the realization that not all the students had the same accessibility to technology from their homes. It is beyond the scope of this paper to discuss accessibility and other obstacles due to technology that raised debates linked to fair treatment, equality and quality of life for the disabled. The main goal of this paper is to discuss how the instructor cultivated an empathic learning environment using the NSEE's eight principles. When the classes were conducted via Zoom, students were provided more opportunities to communicate with the client and his team members on a frequent basis. Extra training sessions were conducted when students were not familiar with the concepts. The instructor continued to encourage the students to share some of dilemmas they had faced during the challenging times.

Reflection: This reflective process is integral to all phases of experiential learning, from identifying intention and choosing the experience, to considering preconceptions and observing how they change as the experience unfolds. Reflection is also an essential tool for adjusting the experience and measuring outcomes. The curriculum was also redesigned to provide opportunities for enhancing their core skills along with technical skills. Through these activities, designed for a face-to-face class, the student learning outcomes were achieved where all the students effectively applied their skills to different projects. This reflective process became an integral to all phases of experiential learning during this new "normal" as the instructor had to reflect back on project's intention and choice of experience, to taking-into-account preconceptions and most important observing how they change as the experience unfolds during the challenging times. With the COVID-19 situation, the instructor decided to modified the assignments and requirements. Further the project of six students was divided into sub-groups of two students where students conducted research, tested, and documented DATIM. This was clearly reflected in the students' assignments, documents, through observations, and their discussions on Zoom.

Orientation and Training: To ensure students receive a full value of the experience of a capstone real business project the client (co-author) had designed a few training sessions about DATIM. This orientation helped the students to understand the background of the project. After the class modality changed in the middle of the semester, more training opportunities were provided. All the students were required to participate in the training sessions on DATIM via a video conference. The intent was that this extra training would provide some extra resource to the students who were struggling with the basic requirements of 508 evaluation of DATIM.

Monitoring and Continuous Improvement: It is important to create learning activities that are dynamic and linked with the student learning outcome. Consequently, the students involved all bore responsibility for ensuring that the experience, as it progressed, continued to provide the richest learning experience possible. Assignments were built into the curriculum that monitored each sub-groups' progress, their challenges, and how they overcame them. For example, group project members on a weekly basis were required to answer the questions in the build-in template: 1) Is your project on schedule? (If not, what and how will you make adjustment(s) to meet required timelines?); and 2) How have you dealt with the new issues? (e.g., what solutions/work-around have you explored or adopted?). In the weekly report, a few more questions were added to specifically address the COVID-19 situation. For example, the following questions were added: "Do any of the changes you have made significantly impact the intent, timeline, or results of your project? If so, please explain", and "Have there been any fortunate by-products or unexpected successes". The last two questions resulted in

interesting observations about projects. The by-products observed by the students in their subgroups turned into lessons learned and suggestions that now are being taken into account by the client's team and the instructor while modifying the DATIM application and teaching the class as remote synchronized via Zoom.

Assessment and Evaluation: In addition to the weekly report, the sub-groups were required to provide documentation and present their project as a final exam assessment and on campus conference. The assessments were designed to develop and refine the specific learning goals, while evaluation provided comprehensive data about the experiential process as a whole and weather it met the course intentions. After eight weeks of face-to-face class leading to a remote class and cancelation of the campus conference, some of the assessment and evaluation process was linked to the student's assignment were modified. Balancing the requirements of the class's objectives and cultivating an empathic learning environment for the students was probably was the most challenging for the instructor. The instructor wanted to encourage the deeper thinking through the assignments but also did not just want to re-create the traditional classroom in an online format. Establishing trust and a consistent presence throughout the course was critical during the unforeseen challenges everyone was facing. An informal discussion platform was created on the learning management system. It was an assignment but not graded. Some students participated in conversation and some hesitated initially. The instructor (author) discussed in class and posted articles about project management and impact of COVID-10. These posts involved everything from timely, respectful correspondence to active, but not overwhelming, participation on the discussion platform.

Acknowledgment: Recognition of learning and impact occur throughout the experience by way of the reflective and monitoring processes and through reporting, documentation and sharing of accomplishments. Both formal and informal acknowledgments were designed into this classroom. Acknowledgment section in the report and an informal suggestions and comments about the project during the breakout room on Zoom were provided. While trying to cultivate an empathic learning environment, the instructor acknowledged all the students, their situation, and provided them an opportunity to express their emotions, challenges and frustrations during these difficult times.

LESSONS LEARNED

Project management class's pedagogy was modified based on previous year's findings. The main intent was to provide students a real business setting where they appreciated the depth of responsibility, accountability and skills needed to work in a team. This year global pandemic's resulted in remote teaching. In addition to the various challenges and obstacles, specific to this class, in the beginning of the semester, the students enjoyed the freedom to design their own sub-projects. However, after eight weeks classroom environment was more of frustration and lack of motivation to complete the projects. Although literature suggest that team work and collaboration skills are crucial when preparing students facing global challenges in the work field, the new "normal" of teaching style (via Zoom) fostered a challenging environment where these skills were not as welcomed by the students. Despite the impact of COVID-19, the instructor was able to cultivate an empathic learning pedagogy by revisiting the eight principle of NSEE's framework. Consequently, an informed learning context that fostered students' growth and actualization of potential in real business setting was created. Developing experiential pedagogy also posed some challenges. For example, some students resisted remote classes and having their cameras on in the Zoom meetings. During re-planning and modifying the assignments, the importance of empathic learning tools incorporated into the classroom became apparent. Re-planning focused on creating a pedagogy that could enable experience to become knowledge. At the same time, the instructor used NSEE's principles to cultivate an empathic learning

pedagogy. As mentioned empathy may be defined as “the action of understanding, being aware of, being sensitive to, and vicariously experiencing the feelings, thoughts, and experience of another…”.

CONCLUDING REMARKS

This paper reflects on the continuing collaborative research which is exploring experiential learning pedagogy for senior information systems students in project management. It specifically reviews the modification made by the author (instructor) in the middle of the semester, when classes changed from in person to remote teaching due to COVID-19 pandemic.

It also discussed how changes in assignments, responsibilities, and presentations were modified during the remote teaching period. This paper discusses how the instructor enhanced the National Society for Experiential Education (NSEE) existing eight principles by including empathy as an additional aspect to the existing principles. The benefits of using NSEE as a foundation to design the capstone project class prior to COVID-19 has proven to be beneficial in providing students an experiential learning environment (Kesar, 2016, Kesar and Pollard, 2020).

KEYWORDS: Project management, Information systems, Experiential learning, DATIM, NSEE.

REFERENCES

Al-Samarrai, S., Gangwar, M. and Gala, P. (2020). The Impact of the COVID-19 Pandemic on Education Financing, World Bank, Washington, DC.

Aucejo, E, French, J., Ugalde, M., & Basit Zafar (2020). NBER, The Impact of COVID-19 on Student Experiences and Expectations: Evidence from a Survey. https://www.nber.org/papers/w27392

Barrett-Lennard, G.T. (1962). Dimensions of therapist response as causal factors in therapeutic change. *Psychological Monographs,* 76(43), 1-13. https://doi.org/10.1037/h0093918

Holm, U. (1996). The affect reading scale: A method of measuring prerequisites for empathy. *Scandinavian Journal of Educational Research*, 40(3), 239-253.

Illanes, P., Law, J., Mendy, A, Sanghvi, S. and Sarakatsannis, J. (2020). Coronavirus and the campus: How can US higher education organize to respond? McKinsey& Company. Retrieved from https://www.mckinsey.com/industries/public-and-social-sector/our-insights/coronavirus-and-the-campus-how-can-us-higher-education-organize-to-respond

Kagan, N. and Schneider, J. (1987). Towards the measurement of affective sensitivity. *Journal of Counselling and Development*, 65(9), 459-464.

Kesar, S. and Pollard, J. (2020). Project Management: Experiential Learning Pedagogy”. In *Societal Challenges in the Smart Society*, Oliva, M., Borondo, J., Murata, K., and Palma, A., Universidad de La Rioja, Spain. Pg. 147- 151.ISBN: 978-84-09-20273-7

Kesar, S., (2016). Including Teaching Ethics into Pedagogy: Preparing Information Systems Students to Meet Global Challenges of Real Business Settings, S. Kesar. *ACM SIGCAS Computers and Society*-Special Issue on Ethicomp, 45(3). https://doi.org/10.1145/2874239.2874303

Loreman T. (2011) Kindness and Empathy in Pedagogy. In: *Love as Pedagogy*. Sense Publishers. https://doi.org/10.1007/978-94-6091-484-3_2

NSEE (2011). https://www.nsee.org/2011-annual-conference

Paradigm Education Solutions (2021). Why Empathetic Teaching Matters: Going the Extra Mile, Online Webinar.

Schleicher, A., (2020). The Impact of COVID-19 on Education: Insights from Education from a Glance, OECD.

Sarkadi and Casmana, A. (2020). The Application of Empathetic Learning in Facing the Covid-19 Pandemic as the Responsibility of Good Citizens., *International Journal of Psychosocial Rehabilitation*, 24(09), 1039-1047. http://doi.org/10.13140/RG.2.2.24178.73923

Ravitch, S. (2020). Why Teaching Through Crisis Requires a Radical New Mindset Introducing Flux Pedagogy.

THE ORIENTATION TO ONENESS OF TECHNOLOGY AND MEANINGS OF LIFE BY PEOPLE IN JAPANESE TECHNOLOGICAI ENVIRONMENTS

Makoto Nakada

University of Tsukuba (Japan)

nakadamakoto@msd.biglobe.ne.jp

ABSTRACT

The purpose of this paper is to clarify the factors which seem to effect or reflect people's views on various matters they face in the informatized environments such as 'their encounter with robots,' the problem of 'autonomous driving car' and others. The author wants to do this by focusing on two points. 1) What kind of values or interest does lie behind these views? 2) How are these values or the interest 'structured' or 'interrelated'? In the past researches that the author has performed in Japan, the author has found that people's views on various problems in the informatized society are strongly or fairly strongly correlated with the orientation to a 'good and virtuous society or life' in their mind. This clearly shows us that some sort of techno-determinism is not adequate for the purpose of understanding people's consciousness in Japanese society. In this paper, the author will continue to examine the past findings using the new data and the author will do another effort with regard to this point by looking at the structure of a kind of an 'shared map of interest' which seems to 'distribute' a set of interest (and ways of understanding) within people's mind.

INTRODUCTION

In this paper, the author will try to make an attempt to enlarge the scope of ethical, cultural and existential discussions on the meanings of people's encountering with the technological matters including robots, AI and others in the informatized environments. The author will do this by focusing on the Japanese data which are gained by the author's own researches performed in Japan for a decade. Through the quantitative analysis on these researches grounded on a kind of qualitative and critical inquiry on the meanings of people's life with the presuppositions that people can't live without various sorts of orientation to the wholeness of meanings, the author has found a very important finding which almost anyone else has never gained in the form of empirical data, i.e. the finding that (at least not a few) Japanese people of today share a kind of ways of thinking and feeling about 'what is a good and virtuous life?' And the author has found too that these ways to pursue the meanings of life tend to determine the direction of evaluation of the meanings of technologies and informatized styles of life. This finding shows us that the technologies don't influence the meanings of life at least in some aspects of life, but rather the meanings of life influence the meanings of technologies. This is quite the opposite direction of influence which a lot of people with orientation to a kind of techno-determinism might expect. For example, so far as the author's research performed in Japan in 2020 shows, people's orientation to a good and virtuous life is found to be strongly correlated with the evaluation of ethical problems in the informatized environments such as the so-called the trolley problem, i.e. the degree of acceptance of the ideas about the choice of victims by the machine. More concretely, people's orientation to the depth of meanings of life to be gained through sharing hardship or sincerity is found to be correlated with people's ways of evaluating the meanings of robots or other technological products in their life.

In the author's view, this suggests us the importance of reconsideration of the meanings of our life in the informatized environments. And in this sense this suggest us that we need to reflect on the meaning of our life by re-examining of various discussions by various authors in Japan and Western cultures with orientation to the wholeness of life in the informaized environments such as Andrew Feenberg, Kitaro Nishida, Rafael Capurro, Bruno Latour, Toru Nishigaki and others.

One of the important points about the wholeness of meanings in life, suggested by our empirical research(s), is that the wholeness seems to be related with a kind of self-reference in Japanese minds, i.e. the tendency that in many cases the experience of something in every-day's life is associated with its self-reflective evaluation in the form of the inquiry, 'what does this mean for our life?' at least in a potential way within one's mind.

And another important point about this wholeness of life and the related consciousness of self-reference is that these seem to reflect some sort of structure of a 'shared map of interest' (or a shared schema of interest (or awareness, attention)) (this is a tentative term).

These consciousness and interest (in people's minds) seem to reflect what can be called a 'shared map of interest.' Specifically, it becomes like this. The analysis of our surveys revealed that people's interest in social issues and issues in social life in Japanese society is structured. In other words, the author has extracted these 5 main factors from the data (2020 research) such as 'integration with nature,' 'orientation to "inner heart" of technology (the potentiality to improve and enrich meanings of life),' 'sympathy for victims (and orientation to sharing inner heart among people),' 'orientation to a society where we live together,' and 'criticism of technology.' These 5 factors were extracted as a result of factor analysis, and it became clear that they constitute what can be called a 'general interest diagram.'

Interests and consciousness in social matters, including the issue of 'human relationship with robots,' seem to emerge in people's mind in the form of 'distribution' of these five interests. At least, within the scope of my analysis, there are 5 interests in almost all social issues (thus the characteristics of these issues emerge in people's consciousness in the form of 'internal interest'). The meanings of these issues consist of a combination of these interests. And the inner interest is the result of the 5 distributed interests. To put this in a different way, the content of the concrete 'meanings' of a particular consciousness of a certain matter or problem can be understood by focusing on the five factors. This seems to be a very important finding for understanding the Japanese mind's inner structure.

Our consciousness is more structured and subdivided than expected. The interrelationship of the areas within this structured consciousness creates the breadth and depth of consciousness and interest. In that sense, Japanese self-consciousness, the orientation to oneness (of the subject and the object as well as the things and the minds) and a sense of unity with nature are more 'modernized' than imagined.

This self-reflective consciousness seems to be a 'practical' one in the sense that individuals can perform reflexivity in the form of emergence of some sort of awareness and cognition in their mind. This point is suggested by various authors. Kitaro Nishida's idea of fusion of thought and practice, i.e. the idea that our experience is originally such a fusion ('pure experience' as the term by Kitaro Nishida), is related to this. And Hideo Kobayashi says, based on the idea of Motoori Norinaga,, *'Mononoaware* (aesthetic consciousness associated with a sense of transience)' is not a noun but a practice to know the origin of *Mononoaware* and Japanese sensitivity itself.'

And in addition to this point, we have to take into consideraton another point, i.e. the point that people's mind and also this reflective and introspective consicousness are 'structued' following the 'shared map of interest' mentioned above. We will examine these points in detail in this article.

ANALYSIS ON PEOPLE'S REFLECTIVE MIND IN JAPAN IN THE INFORMATIZED ENVIRONMETNS

In this section, we want to examine the inner structures of people's minds in Japanese society in the informatized environements by focusing on our research data which are gained by the author's own surveys, in particular by the survey peroformed in Japan in 2020.

The following table, i.e. Table 1 shows the important findings gained through analysis on the data of 2020HTKG Research in Japan.

(The following table shows part of the research findings through the research '2020HTKG.' This is the research done in Japan for 600 respondents in the age of 25-44 living in Fukushima, Miyagi and Iwate Prefectures and also in Chiba and Ibaraki Prefectures. This survey was designed as quota sampling, and ratios of gender and age were quoted from the official statistical report of the Japanese government about Internet users in Japan.)

Generally speaking, the findings of 2020 research are consistent with the findings of the author's past researches with regard to Japanese orientation to a wholeness of life. This time the self-reflection is one of the main foci.)

One of the most important points shown in Table 1 is that people's evaluation or interpretation of the 'potential' meanings of robots, AI or autonomous self-driving cars seems to reflect people's orientation to a 'better and ethical life' rather than mere orientation to 'productivity' or 'efficiency' as indicators of 'success' in competitive and techno-deterministic societies.

Of course, it cannot be denied that people's consciousness can have a different content. It goes without saying that the contents of the survey items and the design of surveys will be reflected in the survey results. However, our goal was to point out the problems of the deterministic schema. In that sense, the content shown in this table is definitive. People's consciousness toward the problems of robots and artificial intelligence is closely related to the deeper 'orientation to the meaning of (better) life." In other words, the problems of robots and artificial intelligence are positioned in the consciousness related to such an 'orientation to creation of the meaning of better life,' and can be understood and interpreted by this orientation.

In short, this suggests that the reflective and introspective consicousness or awareness tend to strongly define the meaning of people's encounters with robots and artificial intelligence. And this suggests us too that the meanings of people's encounter with robots, AI or autonomous cars can't be understood without positinoning these techonoligical issues on a map of meanings of life.

Perhaps by linking this point with the (potential) reflective ideas discussed by Feenberg,Varela and others, we would be able to expand the framework of the discussions beyond Japanese issues.

As can be seen in Table 1, some sort of recursive and subjective consciousness such as 'awareness of the meaning of work' strongly defines the meaning of 'encounter with a robot.' (It is hard to imagine the opposite direction in which robots define recursive consciousness.) However, as shown in the table, concering mere 'length of Internet use' and 'individualist tendency, ' it seems that these have little to do with or have only a weak association with the evaluation of robots in life. On the other hand, the evaluation on the role of the Internet is found to be related with the evaluation of robots and autonomous cars. This suggest us that reflection on meanings of the Internet works in a different way than the mere use of the Internet.

What we can see through this table is that the meanings or evaluations of people's encounter with robots, AI or autonomous cars tend to be formed through the inner self-reflective schema. And this schema is a set of people's self-awareness which seems to be deriving from their cultural-historical-hermeneutical experiences or practices.

Table 1. Correlations between people's views on robots and their orientation to 'a good, virtuous and aesthetic life' (Data: 2020HTKG in Japan) (N=600).

	Denial of autonomous car's judgment for safety including judgment on human life	(The degree of acceptance of the view:) To give a name to a robot will affect human emotion.	Anti-empathy for the trolley problem	Acceptance of robot's diagnosis on condition with support by human doctors	Rejection of robot's diagnosis for one's own children
When you pick up a carefully made product such as a watch, toy, or tableware, you can feel something like the heart of the person who made it. I think that Japanese people generally have that kind of feeling.	**.357****	**.483****	**.408****	**.471****	**.317****
When I hear the story that Japanese interplanetary spacecraft (planetary probe) called Hayabusa which returned to Earth after many years of struggle, I want to say 'thank you, good job' to Hayabusa, even if I know Hayabusa is not a person.	**.309****	**.440****	**.349****	**.432****	**.292****
I do not want to my children to have a hard time to live, but at the same time I hope them to have a lot of good experience through hardship and become a good person.	**.360****	**.427****	**.319****	**.362****	**.373****
The degree of sympathy with *Mononoaware* as Japanese sensitivity with beauty through transience.	**.287****	**.413****	**.367****	**.347****	**.347****
Length of use of the Internet per day	.011	.051	.036	.027	**.185****
Length of time spent on the Internet via smartphone per day	**-.135****	**-.117****	**-.117****	**-.148****	-.069
Evaluation of the importance of the Internet as a source of information	**.247****	**.236****	**.307****	**.276****	**.121****
Individualism	**.137****	**.111****	**.114****	**.083***	**.240****

1) This table shows the correlation coefficients among various matters and views which emerge in people's mind in Japanese society. More concretely, these data show 'how people's orientation to a better and virtuous way of life in the society in the informatized environments reflect or are interrelated with people's views on robots or AI?' 2) The figures of the table show correlation coefficients. 3) **=p<0.01, *=p<0.05, without ** or *=ns= non (statistically) significant. 4) As we can see clearly in this table, people's views on robots or AI are found to be strongly or fairly strongly correlated with people's views on a better and virtuous life. But mere length of use of the Internet has almost no relations (or only limited relations) with the meanings of robots or AI in our society. And 'individualism' is found to be far weaker relations with the meanings of robots or AI in the society. 5) The selection of these items and views shown in this table is due to the author's previous researches as well as the suggestions by the authors questioning the techno-determinism in various ways.

ANALYSIS ON A 'SHARED MAP (SCHEMA) OF INTEREST' IN JAPANESE SOCIETY AS A FACTOR TO DIRECT PEOPLE'S INTEREST AND CONSCIOUSNESS IN THE INFORMATIZED ENVIRONMENTS

In a way, these explanations mentioned above are a continuation of the author's ideas about the potential role of some sort of 'Japanese views of life' in the informatized environments. The author himself believes that it is very important to emphasize this point repeatedly. This is because in today's Japanese society, it seems that a set of 'technical deterministic' ideas and 'competitive principles' values still dominates Japanese society as 'dominant ideas.'

Currently, 'digitization' is being promoted by the government in various areas of society, but it is not explained why it is necessary. Technical determinism dominates Japanese society as if it were an undeniable 'idealism.' Moreover, in conjunction with this, the structure of the universities is about to be transformed into a structure 'suitable for a digital society or a competitive society.' Only 'objective data' that can be processed numerically are emphasized there.

However, it is not understood that, for example, an index of 'productivity' that looks like 'objective' can have various meanings and interpretations. If a company's sales decrease, the 'productivity' and 'production value' per employee will naturally decrease. However, it is also strongly linked to the consumption tendency of society as a whole. In a society where people have no hope for the future, consumption will decrease. Therefore, the sales of the company go down and the productivity also decreases. The existence of such 'linkages' of meanings cannot be understood by simple numerical analysis. Against this kind of Japanese background, the author has repeatedly pointed out the problem of 'technical determinism.' This paper is a continuation of this effort.

And in this paper, the author believes that we can add an additional point to our previous discussions to redirect people's consciousness on the 'meanings' of 'information society.' (As a matter of fact, at least on a latent level, people's consciousness on the 'meanings' of 'information society is found to be directed by their orientation to a better and virtuous life as we have already examined at least partly.) More concretely, it seems that the content of Table 2 shows clearly this new point to be examined carefully, i.e. the 'structured' state of the inner minds.

Table 2. Factors as a kind of schema (a 'shared map of interest') to redistribute an overall interest to each item (Data: 2020 HTKG in Japan) (N=400).

	Orientation to nature, purity with criticism of loss of 'natural' life (Factor 1)	Orientation to social improvement from 'inside' (through 'soul' or 'potential capability to link humans and productivity') of products or tools (or through instrumentality or convenience of tools) (Factor 2)	Orientation to social sympathy through evaluation of victims or sharing hardship (as well as sharing inner emotion) (Factor 3)	Orientation to social improvement through sharing interest in society to live together (Factor 4)	Criticism of (un-criticized acceptance) of robots technology (and other tools) (Factor 5)
Mononoaware	**.187**	.147	**.643**	**.248**	.106
Just as there is a ritual called a needle memorial service, robots and computers that are no longer in use may be offered memorial services.	.081	**.584**	**.198**	.183	**.333**

I am impressed to learn that the last episode of Astro Boy was his self-sacrifice to save the earth.	.099	**.209**	**.672**	**.166**	.081
When I hear the story of someone who helped others at the expense of oneself in a disaster, I also want to rethink of meanings of my life.	**.214**	.038	**.644**	**.362**	.030
When I see a bouquet being offered at the scene of a traffic accident or incident, the images of the victim and the victim's family are clearly in my mind.	.100	**.167**	**.796**	.079	**.192**
It seems convenient to leave the care to the robot, but at the same time, there is a problem because it strengthens the social isolation of the patient.	.163	.192	**.263**	**.228**	**.580**
Just as lifeless earths, mountains and rivers are objects of compassion and empathy, robots will be objects of compassion and empathy in the future.	.055	**.726**	**.179**	.113	**.157**
To prevent the robot from being abused, it is important to have the robot say 'it hurts' and have an emotional expression function.	**.233**	**.550**	.227	.080	**.296**
It's good to use robots for children's education at school to improve learning.	**.221**	**.698**	.080	**.261**	-.054
It's good to use robots on the battlefield so that there are fewer human casualties.	-.003	**.516**	**.185**	-.078	**.330**
It seems that autonomous driving robots using artificial intelligence prevent human mistakes, etc., so even if it is not fully automatic driving, it is okay if safety increases.	**.236**	**.648**	.058	**.312**	-.086

Autonomous driving robots with artificial intelligence seem to be convenient, but there is a problem with easy use, considering that machines are left to make decisions about life and death.	**.323**	.110	.049	**.308**	**.635**
In modern life, humans are too far from nature.	**.656**	.095	**.136**	.120	**.197**
Humans tend to fall (lose purity) when they become too rich.	**.690**	.092	**.125**	.120	**.147**
People have a certain destiny, no matter what form it takes.	**.696**	.076	**.120**	**.146**	.095
In our world, there are many things that cannot be explained by science.	**.728**	.045	.080	**.179**	.086
There are too many people in Japan today who are concerned only with themselves.	**.727**	.075	.092	**.113**	.046
Doing your best for other people is good for you.	**.486**	.100	**.342**	**.302**	.007
The occurrence of big natural disasters is a kind of warning from heaven to humans.	**.537**	.016	**.353**	.055	**.321**
When arranging a job, it is a normal and humane way of thinking that if circumstances permit, you want to prioritize the employment of a close friend over someone who is useful to the company.	.097	**.303**	**.349**	.194	**.238**
When you pick up a carefully made product such as a watch, toy, or tableware, you can feel something like the heart of the person who made it. I think that Japanese people generally have that kind of feeling.	**.212**	.176	**.424**	**.536**	.013

Even if the work at work is difficult, I think that Japanese people generally cannot leave the workplace easily, considering the colleagues who are struggling with them.	**.211**	.158	**.391**	**.437**	.082
(Interest in) Reduction of traffic accidents due to the spread of self-driving cars	**.176**	**.316**	.143	**.676**	-.039
(Interest in) To realize a welfare society like Scandinavia by raising the future consumption tax rate to the extent possible	.059	.119	**.205**	**.634**	**.142**
Response to the question: Are you interested in domestic politics?	.127	.084	**.155**	**.691**	**.256**
Response to the question: Are you interested in global environmental issues?	**.134**	.073	**.210**	**.778**	.075

1)This table shows the result of factor analysis (principal factor method, Varimax rotation). The figures show 'factor loading' (quantitative index showing the strength of 'relationship' of each item with each factor). In this table, in order to interpret the 'characteristics' of each item, we try to show 'how each item has stronger relations with which factor' using the emphasis shown in bold. (Up to 3 of the 5 factors are selected in terms of strength. However, those with a small factor loading value are not emphasized in bold.) 2) The selection of these items and views shown in this table is due to the author's previous researches as well as the suggestions by the authors questioning the techno-determinism in various ways. 3) We used 400 respondents who live in Fukushima, Miyagi, Iwate Prefectures in order to compare these findings in Table 2 with our previous research findings.

As mentioned above, it seems that Japanese minds are more 'structured' than we usually imagine. Table 2 seems to point out to this new way of viewing Japanese mind and Japanese society. We have already found that some sort of views shown in Table 1, which seems to reflect Japanese mind(s), has some kind of strong inner consistency. This can become 'visible' through a kind of statistical methods such as factor analysis. In fact, in the case of 2014 survey (the respondents : 600 men and women in the ages of 25-44 living in Fukushima, Iwate, and Miyagi prefectures), the author could extract one factor, using the data showing 'how people respond to some items which seem to reflect their orientation to a good and virtuous way of life?', namely the data on people's degree of sympathy with these views: 'view of one's fate,' '*Mononoaware*,' 'awareness of compassion for the earth, the earth, mountains and rivers,' 'sympathy for needle memorial services and robot memorial services,' 'sympathy for social victims' and 'respect for privacy awareness and importance of conversation between friends.'

And we have found that this factor (named tentatively as 'self-reflective consciousness') has strong correlations with interest in (or concern for) the following items. 'The damage from the nuclear power plant spreads and threatens the health of children and the younger generation.' 'Family members and loved ones get sick or encounter dangerous situations.' 'I am impressed to learn that the last episode of Astro Boy was his self-sacrifice to save the Earth.' 'Even if the robot has no life, it is reluctant to break it unnecessarily, considering that the robot has the heart of the person who made it.' 'When a house I live in is destroyed due to a natural disaster, I feel like I've lost an important part of myself,

and I feel sad.' This suggests us that people's minds are in the form of a 'structured map' and also this map includes people's interest in various problems. Here is an inner consistency of mind and interest.

And in the case of the 2020 survey, in addition to the items used in the factor analysis of the 2014 survey (or similar items), the following items were added to perform the factor analysis. 'Items related to robot ethics and ethical awareness of self-driving cars' (as shown in Table 1, it is clear that many of them are related to 'orientation to a better way of life'), 'items related to "importance of solidarity consciousness through sharing hardships in the workplace",' 'empathy for craftsmen's work' and others (i.e. 'orientation to a better way of life' in the 'expanded' form), 'items related to the degree of interest in social issues such as domestic political problems and others.'

Then, this time we have obtained five factors as shown in Table 2. This result seems to indicate 'blurring' and 'flickering' of range of consciousness and interest at first glance. However, a closer look at 'how the factor loading score of the factors emerge' would reveal a presence of a kind of a clear internal link (in an extended form) of these five factors. In other words, for example, the 'internal meaning' of each item used in factor analysis is clearly visible. Taking '*Mononoaware* (sensitivity to beauty with a sense of transience' for instance, we have found that this is linked with 'factor related to internal empathy' (factor 3 in Table 2), 'society to live together' factor (factor 4) as well as 'orientation to the unity with nature and a natural life' factor (factor 1).

This is an intuitive interpretation in a sense, but the interpretation of the 'content' of a factor always includes such a 'subjective interpretation.' However, on the contrary, considering that '*Mononoaware*' is linked to 'robot ethics' etc. (this point is already clear in past research), the interpretation of the contents of such factors is not arbitrary. In other words, our data shows that '*Mononoaware*' is not just a 'sentimental consciousness' but also a 'complex consciousness' linked with 'social interest.' This linkage seems to reflect the distribution of factor loadings in Table 2. In this sense, '*Mononoaware*' is not only a 'sentimental consciousness' but also an 'introspective consciousness' associated with even social criticism.

Similar interpretations seem to be possible in the case of other items. These seem to be within the scope of personal interest at first glance, but in fact, these seem to be supported by such a wide range of orientation.

One of the important points shown in Table 1 and 2 is that people are not just evaluating 'robots,' 'artificial intelligence,' 'self-driving cars,' and 'CMC' according to a vague intuition. Rather, it seems that people evaluate a particular problem 'structurally' according to a 'shared map (schema or diagram) of interest (awareness, attention)' that is composed from (or embraces) multiple perspectives.

And concerning some of 'irrational' views in Japanese mind, we can understand what they 'really' mean through analysis using the findings in these tables. For example, in the case of 'disasters as warnings from heaven,' which seems to be absurd at first glance, we know that this is based on 'orientation to natural life and the sense of crisis about its loss' (factor 1), 'orientation to social reform through technologies' (factor 2) as well as 'orientation to society to live together' (factor 4). In this sense, we can regard this as an 'understandable' view, when we see the interest behind it. Similarly, the meanings of other views seem to become 'visible' through this kind of analysis.

RECONSIDERATION ON JAPANESE THOUGHT OF ONENESS

Our starting point in this article was to seek a direction to get out of the framework of thinking influenced by 'technical determinism.' In this regard, it was our expectation that something with

Japanese ideas and ways of thinking would be effective. This premise is basically the same at this stage of the discussion, but on the other hand, it is true that through the examination of the analytical findings shown in Tables 1 and 2, some issues that are lacking in the conventional discussions have become apparent now. For example, the thought of oneness, an unseparated state of subjectivity and objectiveness, a state of fusion of knowledge, emotion and will, i.e. the idea which Kitaro Nishida suggested as our authentic experience seems to be one case.

According to Andrew Feenberg, Nishida's philosophy is a kind of sincere search for Japanese identification in the modernizing and Westernizing period since the beginning of Meiji Era. And also his idea is a criticism of Western thoughts in the sense that these would work as a restraining condition for 'the life as whole.' Feenberg says: Nishida thought that Western culture and science provided people with a way of seeking for reality through rigid adaptation of logic and reason to the reality or the phenomenon itself. And this allowed people to be free from the shackles of constrained cultures and prejudices. But on the other hand, this is the beginning of the other constraints. And this causes some serious problems such as the loss of importance of 'secondary qualities' as a result of being eliminated by the empirical approach. This refers to the state that the objects, the things and phenomena are reduced to purified and abstracted concepts or conceptual entities in the form of 'sense data' or 'brute facts' which would be separated from the immediate and concrete contents of experiences and the things themselves (Feenberg 1995).

If we consider that 'secondary qualities' of things and matters are directly associated with our experiences, so this means the reduction of our experiences too. In this sense, Nishida's effort in the form of criticism of some aspects of modern, Western thoughts is the one to recover 'pure immediacy' or 'the total experience' in the face of dominance of purified concepts.

The author himself sympathizes with this idea and planned the survey in a way that followed this idea. One of the fundamental hypotheses which are included in the author's previous surveys is as follows. 'The fusion of thought, experience, and the interpreted meanings of the object of experience (nature, society, technology, etc.) and the interpretation of experience concerning these objects' is shared by people in the form of one 'overall view of life.' And this serves as a sort of index which people use in the case of assessing individual experiences and social issues. For example, in the case of 2014 research which we have already talked about, we have found that there is a sort of 'reflective or introspective consciousness' in people's mind. It seems that various social phenomena are made something to be experienced and meaningful through this form of consciousness. At the same time, the content of this consciousness seems to correspond responsively to the contents of the observed matters. As we have examined, in this case, the meanings or the contents of evaluation of each item are nothing but an image of the item or the event reflected in the consciousness of the person who is observing that item.

And in the case of 2020 research, the meanings of matters are (or emerge as) the patterns of the 'distributed interest of an overall interest.' In this way, we can say that here is the relation of the knower and the known, as Nishida suggested this as a Japanese way of thinking and experiencing. But on the other hand, this sort of the relation between the knower and the known is found to be in the form of more 'structured' interest. In this sense, we might be able to say that Nishida's idea about relationship between 'the knower and the known' remains within the framework of a kind of 'Cartesian epistemology' (although he criticises it). Our analysis suggests that the range of work by the 'social interests' that correspond to the work of such 'observing self' might be even broader. With that in mind, it is necessary to reconsider the direction of the consideration raised by Nishida and others.

But this doesn't mean that some sort of Japanese ways of thinking and feeling is not 'useful' any more. But on the other hand, as we have already mentioned, through our analysis using the 'shared map of interest' or an 'overall interest schema,' we can confirm that some kind of Japanese activity such as

'give a name to a robot and treat it gently' is not just an activity of animism. According to the figures to show the correlation coefficients between 'the act of giving a name to a robot' and 'the five factors to be shown in Table 1,' this act of naming a robot is found to be strongly correlated with all of these factors (262**(factor 1),370**(factor 2), 217**(factor 3), 427**(factor 4), 174**(factor 5)). This means that this act of naming a robot reflects people's 'deep' and 'internalized' interest in social matters in terms of use of a robot in social environments. And we can imagine that 'naming a robot' is a kind of act that activates the work of these five interests by the act itself and the expressiveness related to the act. In Japanese, we have an expression, '*Mi wo nori dasu*'(leaning out = taking a physical and mental attitude that shows interest in something).This is another example of the duality of an act in the sense that this act or this posture is an act itself and also an expressiveness. This duality would make a kind of situation emerge. In that situation, the target of act, the posture and the interest seem to be interrelated with one another (concerning this point, see: Ichikawa 1992:102).

CONCLUDING REMARKS

While proceeding with this analysis, the author noticed various things, and one important point is why Japanese society is so in the state of 'stagnation,' even though there is so much interest in various issues as suggested by the figures in Table 1 and 2.

Concerning this matter, there is one important idea that we have discovered through our analysis shown in this article. The point is that even for the same problem, the way in which people react is almost completely different depending on 'in what kind of frame the problem would be placed.' For example, the issue of 'cashless society' may be asked in different forms such as 'How interested are you in the following issues?' or 'The following list includes various problems that Japan should currently deal with. Which of these problems do you consider to be of particular importance?' The author used these forms in our questionnaire for 2020 research. And the author has found that the answers are almost completely different according to these forms of questions. The level of awareness of and interest in an individual problem tend to increase as it goes through 'personal interest' (which was also 'social interest' at the same time as we have discussed), even for less familiar problems.

However, in the case of a simple numerical evaluation of importance (not related to one's own 'internal interest,' but just in a form of a kind of mere knowledge or judgement), the degree of awareness of and interest in the problem drops dramatically. In other words, it is no longer 'my problem.'

Perhaps there is a big problem with Japanese society here. As you can see in Tables 1 and 2, people's interests in various problems are higher than expected. Moreover, their interests are interrelated. In that sense, the level of awareness and interest is high. On the other hand, when they are 'out of interest,' the level of recognition drops suddenly and surprisingly in many cases. The difference in the questioning method alone seems to cause such a big change.

Concerning Japanese society, some authors often talk about the duality of 'inside (*Uchi*)' and 'outside(*Soto*).' In Japanese society, people are interested in *Uchi* but not *Soto*. This might be understood: *Uchi* is the sphere where 'meanings' and 'interests' would be distributed to various problems. And this would be done through a distribution of interest through 'the shared map of interest.' In this way, as we have analysed, some sort of 'interest' is distributed to 'sharing of hardship' and to 'autonomous driving cars with a program of self-judgement on life and death.' In other words, the same problem can emerge as a different one as an 'inside' problem or an 'outside' problem at the same time by this kind of distribution of interest.

This is a half-confirmed fact and half imaginative at this moment. By re-analysing the contents of our previous surveys in a more detailed way, we need to re-examine this point carefully. This is the subject for our next step.

KEYWORDS: Japanese thoughts, robots, autonomous car, the wholeness of life, self-reflection, schema of interest.

REFERENCES

Capurro, R., Eldred, M. and Nagel, R. (2013). *Digital Whoness: Identity, Privacy and Freedom in the Cyberworld*. New Jersey: Transaction Books.

Capurro, Rafael (2018). Digitalisierung in der Medizin: Skepsis gegenüber Hypes. *Deutsches Ärzteblatt*, Jg. 115, Heft 31-32, 6: 1426-1429.

Feenberg, A. (1995). The Problem of Modernity in the Philosophy of Nishida. In J. Heisig and J. Maraldo (Eds.), *Rude Awakenings: Zen, the Kyoto School and the Question of Nationalism*. Honolulu: University of Hawaii, pp. 151-173.

Feenberg, A. (1999). *QuestioningTechnology*. New York: Routledge.

Ichikawa, H. (1992). *Seishin to shite no shintai* (Body as Mind). Tokyo: Kodansha.

Introna, L. D. (2007). Maintaining the reversibility of foldings: Making the ethics (politics) of information technology visible. *Ethics and Information Technology*, 9: 11-25.

Kobayashi, Hideo (1979). *Motoori Norinaga*. Tokyo: Shintyosha.

Kimura, Bin (2000). *Guzensei no seisinn byouri* (Psychopathology of coincidence). Tokyo: Iwanami.

Nakada, Makoto (2020). Rediscovery of an existential-cultural-ethical horizon to understand the meanings of robots, AI and autonomous cars we encounter in the life in the information era in Japan, Southeast Asia and the 'West', in Mario Arias-Oliva, Jorge Pelegrin-Borondo, Kiyoshi Murata and Ana Maria Lara Palma(eds.), *Societal Challenges in the Smart Society*, Universidad de la Rioja, Logrono, Spain, 407-418.

Nakada, M. and Capurro, R. (2013). An intercultural dialogue on roboethics. In Nakada. M. and Capurro, R. (Eds.) *The Quest for Information Ethics and Roboethics in East and West*, vol. 1, 13-22.

Nishida, Kitaro (1950). *Zen no kenkyuu* (Studies on goodness) (the original version was published in 1911). Tokyo: Iwanami.

Nishigaki, Toru (1999). *Kokorono jyouhougaku* (Information study seen from perspectives related to human minds). Tokyo: Chikuma Shobo.

Ogata, Tetsuya (2017). *Deep learning ga robot wo kaeru* (Deep learning will change robots). Tokyo: Nikkan Kougyo Shinbunsya.

Varela, Francisco J., Thompson, Evan and Rosch, Eleanor (1991). *The embodied mind: cognitive science and human experience.* MA: MIT Press.

WHY DISABILITY IDENTITY POLITICS IN ASSISTIVE TECHNOLOGIES RESEARCH IS UNETHICAL

Reuben Kirkham

Monash University (Australia)

reuben.kirkham@monash.edu

ABSTRACT

This paper charts a worrying trend in the academic assistive technology community. Assistive technologies research is intended to create new technologies for disabled people, and thereby increase their access to society and enhance their quality of life. In recent years, the academic assistive technology community has become somewhat distracted by attempts to introduce 'disability identity politics' into this research community. This paper argues that these activities do not serve the interests of disabled people, by reducing the opportunity for them to obtain assistive technologies. This practice also disproportionately penalises the most vulnerable disabled people, making it particularly unethical. Through setting out and evidencing this problem, this paper explains why 'disability identity politics' is unethical in the context of assistive technologies research.

INTRODUCTION

Assistive technologies ("AT") research concerns the development and design of novel technologies for disabled people. The development of new assistive technologies is of paramount importance to the inclusion of disabled people in wider society. Indeed, the UN Convention on the Rights of Persons with Disabilities ("the UN CRPD") recognises this as a means for helping address the unfortunate position many people with disabilities live in (including systematic violations of their human rights) (Hendricks, 2007; Ryan, 2020).

In recent years, there has been a small but vocal group of academics who have turned up at the assistive technologies research community, promoting the use of 'disability identity politics', or 'critical' disability studies. As this paper will argue, such activities are a damaging distraction from the core activity that assistive technologies researchers are supposed to be engaging in. The likely effect is to reduce the likelihood of disabled people obtaining the assistive technologies that they need in a timely fashion. This is a serious problem, especially given the high level of social exclusion that many disabled people face, and the low quality of life that many disabled people can have because of this, including poverty (Joseph Rowntree Foundation, 2019), human rights breaches (Ryan, 2020), and difficulties accessing the wider community (Brown et al., 2017; Equality and Human Rights Commission (UK), 2017).

This paper argues against the use of 'critical'[39] disability studies and identity politics in assistive technologies research. In practice, this issue amounts to two startling claims being advanced by a small group of disabled academics. The first such claim is that because these academics have a disability, their general views (or "*demands*" (Ymous et al., 2020)) on assistive technology research must be

[39] For the purposes of accuracy, I have adopted the approach of placing 'critical' in inverted commas in respect of 'critical' disability studies and similar activities: in truth, 'critical' theory is not about the critical thinking that one would hope for from academic researchers but is instead about Foucauldian postmodernism.

followed by this academic community. The second is the substance of their "*demands*", which include *not* taking a evidence-based, "*scientific*" or "*dispassionate*" approach towards the design of assistive technologies, and that the academic community should instead follow a Foucauldian approach based on 'critical' disability studies. Neither of these propositions serve the interests of most disabled people, but instead a minority of academics who happen to have a disability, but unlike many disabled people, have a relatively comfortable lifestyle.

To make this case, this paper identifies five key ethical concerns with the arguments being advanced by those practising 'disability identity politics' in the assistive technologies community. First, this group of disabled academics is not representative of disabled people at large, especially assistive technologies users. Second, it is observed that their conduct is likely to make disability controversial (again), by rejecting widely recognised settlements (e.g. the human rights model of disability). Third, they are diverting resources away from substantive assistive technologies research. Fourth, they are attempting to prevent disabled people from accessing assistive technologies they are entitled to. Finally, instead of promoting human rights, they are promoting pseudo law, a form of 'not law' which is also potentially harmful for those involved in it. Perhaps the overall lesson is that people who claim to be advocates for equality often do the opposite: instead, they often advance their own personal interests.

WHAT IS ASSISTIVE TECHNOLOGIES RESEARCH AND WHY DOES IT MATTER?

Despite the introduction of disability discrimination law over the past three decades, disabled people are often excluded from society. Even in wealthy westernised countries, around 50% of people with disabilities are not employed at all, and those who are employed are often underemployed (Brown et al., 2017). There is a high rate of poverty amongst disabled people and their families (Joseph Rowntree Foundation, 2019). For example, in the UK, the situation that many disabled people are in has been described by the United Nations as amounting to "*grave [and] systematic violations*" of their human rights.[40] Many groups of disabled people are systematically excluded from society due to the failure to make reasonable adjustments, and thus "*accessing appropriate housing, the built environment, transport, information*" continues to be challenging (Equality and Human Rights Commission, UK, 2017). The result is that some disabled people have a relatively poor quality of life, compared to people without disabilities.

Assistive technologies are of vital importance in helping address these problems. It is of such importance that the UN CRPD – the international human rights treaty for disabled people – requires states "*to promote the design, development, production and distribution of accessible information and communications technologies and systems at an early stage, so that these technologies and systems become accessible at minimum cost*" (per Article 9(h)) and sets out a general obligation to "*undertake or promote research and development of, and to promote the availability and use of new technologies, including information and communications technologies, mobility aids, devices and assistive technologies, suitable for persons with disabilities, giving priority to technologies at an affordable cost*" (Article 4(g)). All this is said in a convention that has been ratified by over 170 nation states, underscoring the relevance of this concern.

Assistive technologies research is therefore funded as a mechanism for a state to meet its international human rights obligations. It arises from the recognition that the full inclusion of disabled people in today's society is not always economically feasible. Realistically, it is only by providing readily available

[40] See the UN Report CRPD/C/15/R.2/Rev.1
(https://www.ohchr.org/Documents/HRBodies/CRPD/CRPD.C.15.R.2.Rev.1-ENG.doc)

technologies that remove barriers, people with disabilities are more likely to be included in society. A considerable proportion of assistive technologies research focusses on the creation of new assistive technologies ("AT"), including 'bespoke' or DIY AT for individuals that responds to individual needs (Ellis et al., 2020; Hook et al., 2014), and creating AT in an economic fashion (de Witte et al., 2018), such as via 3D printing or fabrication (Hamidi et al., 2018; Schwartz et al., 2020). Other work in the AT field focusses on the development of standards so that services are accessible: perhaps the most prominent example of this is the WCAG guidance for web accessibility (Lazar et al., 2016). Another area of important activity is empirical research that examines the use of AT on the ground, with a view towards developing practical design principles for disability inclusion (e.g. Lumsden et al., 2017; Shinohara and Wobbrock, 2011), or the monitoring of compliance (e.g. Lazar et al., 2010; Rutter et al., 2007).

The current position is that AT research has advanced the inclusion of disabled people, but a lot more AT research still needs to be done to achieve genuine disability equality. When considered against the range of disabilities and the demand for AT research, there is a chronic lack of funding. The result is that most work on AT research has so far focussed on a minority of disabled people: for instance, around 50% has focussed on people with vision impairments (Mack et al., 2021), even though this is less than 10% of disabled people.[41] Accordingly, there is a critical need to expand the scope and extent of AT research.

WHAT HAS GONE WRONG IN THE ASSISTIVE TECHNOLOGIES COMMUNITY?

The academic assistive technologies community is really a part of the wider human computer interaction ("HCI") community. It primarily operates through academic conferences, most notably ACM ASSETS (the main assistive technologies conference) and ACM CHI (the largest human computer interaction conference), with other conferences and journals having a lesser role. The largest groups focussed on AT research worldwide are at the University of Washington, Carnegie Mellon University and Monash University: most other AT researchers are in smaller groups, with perhaps one or two academic faculty members focussed on AT research within a wider group of HCI researchers.

The result is the wider HCI community has a considerable influence on AT research. Indeed, it is in the wider HCI community where the issue raised by this paper largely started. Over the course of the past three decades, the HCI community has welcomed social scientists and other academics in humanities departments. Some social scientists have contributed considerably to HCI, providing empirical results and approaches that enable more effective design of interactive systems. Unfortunately, along with them also travelled the 'critical' theorists, who were able to stick around and expand due to an absence of an effective mechanism to remove them. As explained in (Button et al., 2015), an earlier challenge to the inappropriate use of 'critical' theory in HCI was met as follows:

> "Our paper upset people from a number of different groups. … The panellist from IBM played music to us, seeming to suggest that we suspend our intellectual concerns in order to be happy and get on well with other people, and the panellist from academia spoke in defence of those we had criticised, and attempted to show how our criticisms were based upon ill-informed, partial and distorted readings of their work. … With the exception of those whose work we criticised, we were puzzled by the paper's reception. Critique is not new for CHI."

[41] See for example https://www.and.org.au/pages/disability-statistics.html

There is now little in the way of an effective 'battle of ideas' in HCI: once something is out there, HCI tends to get stuck with it. The critical theorists have been able to make their way in incrementally, for instance by claiming to be doing 'Humanistic HCI' and that modified review processes (in practice less rigorous ones) were needed to include these 'new perspectives' (Bardzell and Bardzell, 2015). Today, the 'critical' theorists are so established that they have formed their own 'critical and sustainable computing' subcommittee at ACM CHI, then promptly decided to nearly double the acceptance rate (to 45%) for their own papers. As one commentator dryly observed, this committee is "neither critical nor sustainable".[42]

In the assistive technologies community, the appearance of 'critical' theory is a more recent phenomenon. Asides one paper that raised disability studies in 2010 (Mankoff et al., 2010), this has not explicitly re-appeared until around two years ago[43], with the appearance of an 'experience report' at ACM ASSETS asserting that "*SIGACCESS and SIGCHI are being shaped by new members as they integrate feminist theory, critical race theory, postcolonial studies, and critical disability studies into the field*" (Williams and Boyd, 2019). Since then, there has been a flurry of activity, including a planned workshop on Critical Disability Studies which was to be held at CHI 2020 as well as publishing several papers within the accessibility community (e.g. Hofmann et al., 2020; Williams et al., 2021; Ymous et al., 2020), including one which contains a list of "*demands*" (Ymous et al., 2020).

At the heart of the concern of this paper is these substantive "*demands*" (Ymous et al., 2020). There are two interrelated claims being made:

1. Disabled academics have a special knowledge of disability, so their viewpoints should be given special weight. The fact that they are disabled is more important than the *substance* of what someone has said. As a result, *only* disabled academics should determine how the AT research agenda is shaped, and equally, what is a valid finding depends on *whom* has made the findings (rather than this being determined on an objective basis using empirical evidence).
2. The Assistive Technologies community should *not* follow a scientific or empirical approach towards its work, but instead follow 'critical' disability studies (i.e. postmodernism and the teachings of Foucault and his successors (Pluckrose and Lindsay, 2020). Academics who fail to follow this 'critical' approach are not just wrong in the view of 'critical' disability studies, but are causing "*violence*" to disabled people (Williams and Boyd, 2019; Ymous et al., 2020).

These claims have multiple problems with them. Whilst the 'critical' practitioners seek to morally impugn anyone who is not religiously following their agenda, from an *objective* ethical standpoint, it is their approach which is unethical. To concrete this, I now provide five specific reasons why they work against the interests of disabled people at large and how they are unethical.

1. Disabled academics do not really represent most disabled people

The identity politics practitioners start with a kernel of truth. The design of assistive technology greatly benefits from the inclusion of disabled people: there is a considerable body of empirical evidence that demonstrates this (e.g. Hurst and Tobias, 2011; Shinohara and Wobbrock, 2011). There are many disabled academics in the assistive technologies community who successfully draw upon their experience of having a disability, both as a driver for their work and in order to design *objectively* more

[42] https://docs.google.com/document/d/1p_dabojzr58_D84VOWJvcdWrYCjd-4P2UEFt-foNo_E/edit

[43] The absence of such work until 2019 has been bemoaned by one of the 'critical' disability studies academics in a recent issue of ACM Interactions (Williams et al., 2021).

effective assistive technologies. When this works well, then this is a positive step towards advancing assistive technologies research. However, the identity politics practitioners take this further, claiming that as they are academics with a disability who *assert* they are assistive technologies researchers, their *opinions* should carry special weight. In effect, they claim to be proxies who should be consulted (and their "*demands*" followed), in lieu of consulting disabled people at large, or conducting empirical evaluations of assistive technologies.

The underlying difficulty is that the average *academic* with a disability has very different life experiences than the average *person* with a disability. As noted above, most disabled people, even in Western societies, are unemployed (or underemployed) and are subject to systematic discrimination in their day-to-day lives. This is a world away from the average academic with a disability, whilst likely be suffering from discrimination in their jobs (see e.g. Kirkham et al., 2016; Sukhai and Mohler, 2016), will still normally have had a middle class experience and lifestyle. The identity politics approach has also been criticised for overlooking the importance of impairment in the lives and politics of disabled people (Jenks, 2019), again underscoring the distance between 'critical' disability studies practitioners and most disabled people. The truth is that the needs and goals of these purported 'proxies' is far away from the average disabled person.

Unfortunately, this is reflected in the shape of the research that is done in the assistive technology community – very little has been done on disability benefits (despite it being a major issue (Ryan, 2020), and until recently, people with intellectual disabilities and cognitive impairments were mostly excluded (Mack et al., 2021). Enabling identity politics is only going to exacerbate this problem. Disabled people should be consulted more, but that means a wider consultation, rather than using proxies that are mostly representing themselves. In fact, stronger governance is needed to ensure a fair distribution of AT research across the disability community, and that this support is provided to those most in need.

Enabling disabled academics to have *control* over what research is done would not be the good governance that AT research requires. The unhappy truth is that disabled academics are often in a hopelessly compromised position, especially those without secure academic positions. This is because the academic community systematically fails to make reasonable adjustments: the same point often applies to the universities that employ disabled academics, too. The result is that we have identity politics practitioners "*demanding … accountability*" (in their words), but really their interest is cocooning themselves within a sinecure in which they are unaccountable to those disabled people who sorely need AT. This is not in line with the interests of those disabled people who would need new AT to be being created for them.

This problem can be seen on the ground. When given responsibility to improve the opportunities for disabled academics, this group decided to focus on making conferences accessible for people who are already there, rather than helping those who are not part of the community.[44] The absence of complaint when telepresence attendance was dropped – vital for many disabled academics who are unable to travel (be it due to funding or their disability itself (Kirkham et al., 2016) - is one example of this. Another is the failure to challenge the peer review process, which again is exclusionary of many people with disabilities (Kirkham et al., 2015). Instead of broader change, the focus has been on the

[44] There is a group called Access SIGCHI which overlaps heavily with the 'critical' disability studies practitioners. The problem is apparent from their reports (e.g. Mankoff et al., 2020): their interests are focused on the physical conference being accessible for those who can afford to attend it, rather than being on providing the same experience without travelling, or correcting the peer review process not to be discriminatory.

narrow issue requiring *authors* to make paper *submissions* accessible to screen-reader users[45], which is both a distraction and likely to be discriminatory against other groups of disabled people.[46] The steps at including disabled academics who are outside of the existing community are so ineffectual, they almost appear to be intended to ensure that potential competitors to the existing disabled academics do not get access or inclusion. In other words, this group claims to be including disabled people at large, but have largely focussed on helping themselves.

2. Making disability inclusion controversial

Historically disabled people have been excluded from decision making about them. There remains an unhappy history of disability discrimination being the norm, even in westernised societies. The international standards in respect of inclusion of disabled people are uncontroversial – these are now all agreed by way of the UN CRPD. Nor is it controversial to claim that people with disabilities are widely discriminated against. Similarly, we have moved from the medical model and onto the social model, where people with disabilities are to be included in decisions about them.

It is notable that the academic assistive technologies community has been at the forefront of moving away from a medical model and in other innovations that improve the inclusion of disabled people. To give some examples, this community has been relatively successful in including a range of academics who have disabilities, and generally insists that disabled people are participants for studies (as opposed proxies), helping to ensure that AT design is properly evidenced. Yet, at the same time, there is no requirement for AT research to be published within the AT community: there are other academic outlets in the wider HCI community which also publish such work. This means that the AT community and its institutions have *influence*, rather than *control* over the wider academic AT agenda.

The 'critical' activists seek to make the way in which disabled people are included controversial once again. Outside of the AT community, the 'critical' activists (at large) have already succeeded in making the Social Model subject to doubt (see Pluckrose and Lindsay, 2020): they also risk sweeping up *empirical* disability studies in a similar manner. Such efforts can only be damaging to the interests of disabled people. If the activists within AT community succeed, the likely outcome is they will undermine genuine advances in disability inclusion in AT research. After all, it is easy enough for AT researchers to leave the community and present their work elsewhere: indeed, why would most academics wish to put up with identity politics and all the unpleasantness that goes along with it? The controversy that the identity politics activists seek to bring can only be disruptive, and risks moving the existing AT community out of the mainstream and with it, all the advances it has brought for disability inclusion.

3. Diverting resources

There is a critical need for more assistive technologies research to be done. The result of the present lack of funding is to diminish the quality of life of many disabled people. It is important that the

[45] One irony of this is that a 'critical' disability studies practitioners like to make their papers as difficult to follow and understand as possible (Murray, 2019). This naturally excludes people with a range of cognitive impairments, or indeed those unable to invest a considerable amount of time in attempts to understand it (Berghs et al., 2016).

[46] For the small number of cases, the conference could employ experts manually make the relevant paper submissions accessible for *reviewers*, without relying on authors to do this (in a probably imperfect manner). Moreover, requiring authors with certain disabilities (e.g., relevant specific learning difficulties) to conduct this elaborate exercise at the submission deadline is likely to be discriminatory.

resources available for assistive technologies research are used wisely, with the maximum impact possible for disabled people. Delays in providing assistive technology mean delays in improving the quality of life of disabled people, many of whom need new assistive technologies to be created to have a reasonable quality of life.

Time spent on identity politics and Foucauldian 'analysis' amounts to public resources being diverted away from creating assistive technologies. There is a finite number of jobs for assistive technologies researchers and faculty members. Similarly, there is a relatively limited amount of research funding available specifically for assistive technologies research. If these positions are taken up by practitioners of identity politics seeking to procure a sinecure to practice postmodernism, then this reduces the funding and resources directly available to conduct *bona fide* assistive technologies research.

This issue is not simply about taking away funding from assistive technologies research. The best assistive technologies research is translational: it involves working with others to create innovations that improve the lives of disabled people. Identity politics has an off-putting quality to it: indeed, the rationale of 'critical' theory is to exclude people, by creating a memeplex of expected language and requirements, and 'trip wires' for the unwary (Murray, 2019; Pluckrose and Lindsay, 2020). It is not difficult to see how off-putting this could be for (more generalist) computer scientists outside the assistive technology community: thus, reducing the amount of collaboration (and therefore AT research) that would otherwise take place.

The likely result is that many disabled people will get less of the assistive technologies that they sorely need to have a better quality of life. One imagines that when most disabled people eventually realise what is happening, they will be rather disappointed and upset. The simple truth is that identity politics practitioners are using the fact they have *some* disability (albeit whilst living reasonably good lives) to take away resources from other disabled people. It arguably amounts to a misappropriation of public funds, where funds are ultimately being siphoned away from helping the most vulnerable in our society.

4. Obstructing access to assistive technology on ideological grounds

The identity politics practitioners are not content to just delay the provision of assistive technologies, by distracting from their production. They also wish to decide whether certain assistive technologies should be allowed to be used by other disabled people. In other words, they seek to take important decisions about disabled people's lives away from the individual and arrogate these decisions to themselves. Of course, this is inconsistent with the premise of identity politics and 'listening' to disabled people, as it is an enterprise of identity politics practitioners using their identities as disabled academics to constrain the choices available to *other* disabled people (whose identity-based views presumably don't count, for reasons that have not been explained).

Unfortunately, some advocates of disability identity politics have been claiming that certain *bona fide* assistive technologies should be prohibited (Pluckrose and Lindsay, 2020), even though these technologies might be genuinely beneficial. By way of an example of such a proposal, consider (Williams and Gilbert, 2020), where in response to a paper (Kirkham and Greenhalgh, 2015) saying that some disabled people (with relevant impairments) might have the *right* to use cognitive protheses to interact more effectively in social situations, this is asserted to be wrong because:

> "The Neurodiversity and Disability Rights Movements firmly object to notions that the onus of inclusion lie with the disabled person. … In these moments, consider society's responsibility to

> combat stigma and bigotry and accept embodied difference as a valid and worthy way of being in the world."

In the real world, there are many problems with such a claim. First, the "*Neurodiversity and Disability Rights Movements*" are not a homogenous entity with a unanimous position in such issues, nor do they represent most disabled people. Indeed, the underlying claim is really copied from 'critical' disability 'theory', which given its inaccessible nature, most people with disabilities are simply unaware of (Berghs et al., 2016), let alone having subscribed to. The approach advocated by most mainstream disability advocates is a lot more realistic and embodied by the UN CRPD, namely that disability inclusion is a balancing exercise and reasonable (but not unlimited) steps should be taken. The notion that everyone will simply drop social expectations and be fully inclusive is unfortunately just a utopian fantasy: the unhappy truth is that disabled people are often being increasingly discriminated against (consider for example, the treatment of disability benefits claimants (Ryan, 2020)). Surely someone with a disability should at least have the choice to use technology to escape from such a situation, rather than being condemned to indefinitely wait on *ideological* grounds for full disability inclusion, Micawber-like, to just 'turn up'?

There are other troubling claims being made by the identity politics practitioners in the assistive technologies community. As part of their pursuit of postmodernism, the proponents of 'disability identity politics' have even gone as far as to make claims that rational thought and evidence are inappropriate. For example, it is asserted that "*it is time to move beyond fantasies that our work is apolitical, objective, neutral, or dispassionate. … Ahistorical science produces violence no matter how rigorous or hard the field is" (*Williams and Boyd, 2019), that the scientific method is objectionable because "*statistical principles also prevent inclusivity*" and using "*control group[s] … violates the principle of Justice"*[47]*,* or that "*epistemic injustice starts with questioning the disembodied and dispassionate ways we deem appropriate for knowledge production*" (Ymous et al., 2020). The simple truth is that *objective* evidence is necessary to challenge (genuine) injustices, to demonstrate if an assistive technology is working (or not), or to obtain funding for assistive technologies. In abandoning *objective* evidence, the result is that appropriate assistive technologies are less likely to be provided.

The obstructions being raised here are inherently unethical, and well below the professional standards expected of computing professionals (including AT researchers). For instance, the *ACM Code of Ethics,* imposes professional expectations and a minimum standard of competence, per Part 2 of the Code generally. As such, these activities should be rejected as being inappropriate, both on a substantive harm analysis, but also that they do not align with the explicit standards for professional ethics.

5. Promoting pseudo-law

One problem with discrimination law is that it often does not operate effectively on the ground. The main issue is that it is complaint based, where an individual claimant must bring proceedings in their own right. This is an expensive and onerous undertaking that takes place in public: the result is that many claimants are deterred, and in turn, discrimination law is not properly enforced. Identity politics practitioners have also been following a 'complaint' based approach, where they loudly and angrily object to anything that they disagree with. But those who shout the loudest do not normally reflect

[47] These last two quotations appear in a paper (https://perma.cc/CGW7-LV5U), which was due to be presented at a critical disability studies workshop https://katta.mere.st/nothing-about-us-without-us/ (https://perma.cc/4HJR-B3Z7) at CHI 2020.

those who have the greatest need. This is especially true in respect of disability: it is those who are the most impacted by disability who are least likely to have a voice.

Discrimination law is a balancing act, which involves addressing competing considerations in a fair manner. The difficulty with the approach being adopted by the identify politics practitioners is that they are side-stepping this balancing exercise, and "*demanding*" (Ymous et al., 2020) that their result (copied from 'critical disability studies') be followed. As noted above, these 'results' tend to 'balanced' away from the interests of the most vulnerable members of society and in favour of the interests of them promoting them. Instead, these activities involve considering diversions and irrelevancies raised by 'identity' politics practitioners, whereas in reality these are matters that cannot (and should not) be given any weight at all in a human rights analysis.

The alternatives to the UN CRPD and anti-discrimination law being advanced by identity politics advocates are pseudo-legal in nature. Pseudo-law is a form of fake-law, normally promulgated by tax-protestors, or other people who wish not to be subject to the laws of the land (McRoberts, 2019; Netolitzky, 2018b). Netolitzky describes this phenomena as being a group of people "*who claim extraordinary authority and immunity, [making] claims [that] are purportedly expressions of legal rights and principles [following] an alternative, different set of rules that mimic or ape the structure and language of "conventional" law"* (Netolitzky, 2017). As explained by Justice Rooke in the celebrated case of *Meads v Meads*, "*Mediaeval alchemy is a helpful analogue … Alchemists sold their services based on the theatre of their activities, rather than demonstrated results, or any analytical or systematic methodology … they promise gold, but their methods are principally intended to impress the gullible*".[48] Such activities are, to put it mildly, far from being constructive or appropriate, especially if being practiced by an academic community.

Unfortunately, 'critical' disability studies follows the same notorious model: it is promoted as an alternative form of (discrimination) law which must be followed, whilst often being the opposite of what the law actually requires. Just as with pseudo-law, it involves objecting to rational thought, the application of evidence to facts: instead it operates by citing a range of largely incomprehensible sources and claims that are said to be mandates and requirements. In this case, the pseudo-discrimination-law[49] being advanced operates by preying on the failings of real discrimination law. It presents a superficially attractive approach, by exploiting the fact that discrimination law is not enforced often enough on the ground (and is thus said not to really work at all). Yet the simple fact that something is an *alternative* does not make it better than the current system: whilst discrimination law often helps disabled people with legitimate claims, identity politics practitioners only help one group, namely themselves.

There are potentially very serious consequences for people who engage in pseudo-discrimination-law. Promoting some of the ideas in this area whilst holding out an expert is something that is likely to attract civil liability for unlawful disability discrimination: this is because it would amount to incitement of discrimination (especially attempts to prevent the use of assistive technologies based on Foucauldian objections).[50] In certain circumstances, it can attract criminal liability, such in the UK

[48] Meads v. Meads, 2012 ABQB 571 at [78].

[49] It is somewhat curious that works that challenge 'critical' theory fail to pick up on this connection, as pseudo-discrimination-law is probably the most accurate description of the underlying phenomena. Perhaps this is due to the relative youthfulness of pseudo-law being an area of academic and professional comment (Netolitzky, 2018a).

[50] See for example s.111 of the UK Equality Act (2010).

under Part 8 of the Equality Act (2010) (where the penalty is a fine[51]), or in Australia under s.43 of the Disability Discrimination Act (1992) (potentially carrying a penalty of 6 months imprisonment[52]). Moreover, if there is a misappropriation of public funds – for instance if someone were to take funding for a project on assistive technology, but then siphon some of it into identity politics – then there could also be significant criminal law consequences.[53]

Such an enterprise is unethical: there is no good reason not to follow discrimination law or to reject a widely respected human rights treaty (i.e. the UN CRPD). When the result is to damage the interests of disabled people at large, this is particularly troubling, likewise when what is being presented is a form of pseudo-law. It is also the opposite of the *ACM Code of Ethics*, which expects a proper respect for the rule of law (see 2.3 in the code) and challenges the promotion of unlawful discrimination (see 1.4). The appearance of pseudo-law also illustrates a broader problem: such works only end up being published with weak academic governance in the AT community, raising a wider issue to be addressed. Putting in place measures to ensure that AT research is always rigorous, including in respect of compliance with the ethos of the UN CRPD and discrimination law is a wider goal of importance: addressing this would ensure these public resources are always spent wisely.

CONCLUSION

There are many serious ethical problems raised by the efforts to bring forward 'critical' disability studies and/or identity politics in the assistive technologies community. The main flaw is that this will take resources away from people who desperately need them, and instead spend them on unsupported pontification (including Foucauldian 'analysis'). The delayed provision of new assistive technology is often the *denied* provision of assistive technology and with that, the improved quality of life that disabled people are entitled to.

The problem identified in this article is not unique to disability. In a recent article on race relations in the United Kingdom (Phillips, 2021), the former head of the UK Equality and Human Rights Commission, Trevor Phillips OBE, had this to say:

> "Sewell has tried to bring a scientific approach to a problem that bedevils western societies, and as with all science, new data often means that we need to change our theories. Depressingly, a minority want the debate about race to continue as a medieval contest of faith, in which the catechism — "institutional racism", "white privilege" — is mouthed unthinkingly, without understanding. Those who deviate are lashed as heretics. To my mind it is the self-proclaimed radicals who are, in fact, least keen on change. For the zealots to justify their revolutionary aims, women, disabled people and sexual and ethnic minorities must remain in suffering."

[51] See s.112 of the Act.

[52] The Act provides that "It is an offence for a person: (a) to incite the doing of an act that is unlawful under a provision of Division 1, 2, 2A or 3; or (c) to assist or promote whether by financial assistance or otherwise the doing of such an act." The relevant Divisions prohibit direct or indirect disability discrimination, including the failure to make (legally) reasonable adjustments.

[53] Consider for example the case of Eva Lee, who provided false information to the National Science Foundation and was prosecuted for the provision of this false information – that case has nothing to do with identity politics, but suitably illustrates the risk involved. (https://www.sciencemag.org/news/2020/04/georgia-tech-researcher-pays-high-price-mismanaging-nsf-grant).

The assertions being advanced by 'critical' disability 'scholars' suffer from the same flaws: they reject evidence (and "*demand*" it not be followed), seek to prevent progress, and constitute the recitation of illogical mantras in place of substantive improvements. They are just as harmful, as Phillips says, "*for the zealots to justify their revolutionary aims … disabled people … must remain in suffering*": this is exactly what denial of assistive technologies to those in need of them means on the ground.

There is an active ethical duty to put an end to identity politics and similar activities in assistive technologies research. Allowing such activities is also contrary to state parties obligations under the UN CRPD. To address this, it is possible to retract such works that have been published[54], and to take steps to check that people engaged in such activities are not given public funding for assistive technologies research. There is also the need to prevent a repeat in the future: next time people 'appear' in the assistive technologies community asserting they are advancing equality, these claims should be subject to anxious scrutiny and rigorous analysis.

A discussion on how to address 'critical' disability studies and whether or not this complies with existing ethical codes is of great importance. Ultimately, most of the people behind it say they want the same goal – the inclusion of disabled people in wider society – the problem is that their approach is misguided, often the opposite of what the law requires, and is bound to fail. It is time to adopt a more ethical approach towards the development of assistive technologies that does not rely on identity politics, but instead is based on evidence and human rights. At the same time, we should heed the lesson of this unhappy story: people who say they are advancing equality are often just advancing their own private interests.

KEYWORDS: assistive technologies, disability discrimination, ethics, identity politics.

REFERENCES

Bardzell, J. and Bardzell, S. 2015. Humanistic HCI. *Synthesis Lectures on Human-Centered Informatics*. 8 (4), 1-185.

Berghs, M.J. et al. 2016. Implications for public health research of models and theories of disability: a scoping study and evidence synthesis.

Brown, D. et al. 2017. Tackling gender, disability and ethnicity pay gaps: a progress review. *Equality and Human Rights Commission (EHRC)*.

Button, G. et al. 2015. *Deconstructing Ethnography: Towards a Social Methodology for Ubiquitous Computing and Interactive Systems Design*. Springer.

Ellis, K. et al. 2020. Bespoke Reflections: Creating a One-Handed Braille Keyboard. *Assets 2020*.

Equality and Human Rights Commission (UK) 2017. *Disability rights in the UK: UK Independent Mechanism updated submission to the CRPD Committee* https://www.equalityhumanrights.com/sites/default/files/crpd-shadow-report-august-2017.pdf.

[54] The ACM Policy (https://www.acm.org/publications/policies/retraction-policy) provides that papers can be retracted if there is "clear evidence that findings are unreliable as the result of either errors or misconduct". Pseudo-discrimination-law could easily fall within that policy.

Hamidi, F. et al. 2018. Participatory design of DIY digital assistive technology in Western Kenya. *Proceedings of the Second African Conference for Human Computer Interaction: Thriving Communities*, 1-11.

Hendricks, A. 2007. UN Convention on the Rights of Persons with Disabilities. *Eur. J. Health L.* 14, 273.

Hofmann, M. et al. 2020. Living Disability Theory: Reflections on Access, Research, and Design. *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 1-13.

Hook, J. et al. 2014. A study of the challenges related to DIY assistive technology in the context of children with disabilities. *DIS 2014*, 597-606.

Hurst, A. and Tobias, J. 2011. Empowering individuals with do-it-yourself assistive technology. *ASSETS 2011*, 11-18.

Jenks, A. 2019. Crip theory and the disabled identity: why disability politics needs impairment. *Disability & Society*. 34 (3), 449-469.

Joseph Rowntree Foundation 2019. *Poverty rates in families with a disabled person (https://www.jrf.org.uk/data/poverty-rates-families-disabled-person)*.

Kirkham, R. et al. 2015. Being Reasonable: A Manifesto for Improving the Inclusion of Disabled People in SIGCHI Conferences. *Alt.Chi 2015* (New York, NY, USA, 2015), 601-612.

Kirkham, R. et al. 2016. Using Disability Law to expand Academic Freedom for Disabled Researchers in the United Kingdom. *Journal of Historical Sociology*. 29 (1), 65-91.

Kirkham, R. and Greenhalgh, C. 2015. Social Access vs. Privacy in Wearable Computing. *Pervasive Computing, IEEE*, 14 (1), 26-33.

Lazar, J. et al. 2010. Up in the air: Are airlines following the new DOT rules on equal pricing for people with disabilities when websites are inaccessible? *Government Information Quarterly,* 27 (4), 329-336.

Lazar, J. et al. 2016. Human–Computer Interaction and International Public Policymaking: A Framework for Understanding and Taking Future Actions. *Foundations and Trends® Human–Computer Interaction*. 9 (2), 69-149.

Lumsden, J. et al. 2017. Disabilities: assistive technology design. *Encyclopedia of Computer Science and Technology*, 390.

Mack, K. et al. 2021. What Do We Mean by" Accessibility Research"? A Literature Survey of Accessibility Papers in CHI and ASSETS from 1994 to 2019. *arXiv preprint arXiv:2101.04271*.

Mankoff, J. et al. 2010. Disability studies as a source of critical inquiry for the field of assistive technology. *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*, 3-10.

Mankoff, J. et al. 2020. 2019 Access SIGCHI report. *ACM SIGACCESS Accessibility and Computing*, 126, 1-1.

McRoberts, C. 2019. Tinfoil Hats and Powdered Wigs: Thoughts on Pseudolaw. *Washburn LJ*., 58, 637.

Murray, D. 2019. *The Madness of Crowds: Gender, Race and Identity; The Sunday Times Bestseller*. Bloomsbury Publishing.

Netolitzky, D.J. 2017. Organized Pseudolegal Commercial Arguments as Magic and Ceremony. *Alta. L. Rev.,* 55, 1045.

Netolitzky, D.J. 2018a. After the Hammer: Six Years of Meads v. Meads. *Alta. L. Rev.,* 56, 1167.

Netolitzky, D.J. 2018b. Lawyers and Court Representation of Organized Pseudolegal Commercial Argument [OPCA] Litigants in Canada. *UBCL Rev.,* 51, 419.

Phillips, T. 2021. Silence of white establishment betrays Sewell. *The Times*.

Pluckrose, H. and Lindsay, J.A. 2020. *Cynical Theories: How Activist Scholarship Made Everything about Race, Gender, and Identity—and Why This Harms Everybody*. Pitchstone Publishing (US&CA).

Rutter, R. et al. 2007. *Web accessibility: Web standards and regulatory compliance*. Apress.

Ryan, F. 2020. *Crippled: Austerity and the demonization of disabled people*. Verso.

Schwartz, J.K. et al. 2020. Methodology and feasibility of a 3D printed assistive technology intervention. *Disability and Rehabilitation: Assistive Technology*. 15 (2), 141-147.

Shinohara, K. and Wobbrock, J.O. 2011. In the shadow of misperception: assistive technology use and social interactions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 705-714.

Sukhai, M.A. and Mohler, C.E. 2016. *Creating a Culture of Accessibility in the Sciences*. Academic Press.

Williams, R.M. et al. 2021. Articulations toward a crip HCI. *Interactions,* 28 (3), 28-37.

Williams, R.M. and Boyd, L.E. 2019. Prefigurative politics and passionate witnessing. *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 262-266.

Williams, R.M. and Gilbert, J.E. 2020. Perseverations of the academy: A survey of wearable technologies applied to autism intervention. *International Journal of Human-Computer Studies*, 102485.

de Witte, L. et al. 2018. Assistive technology provision: towards an international framework for assuring availability and accessibility of affordable high-quality assistive technology. *Disability and Rehabilitation: Assistive Technology*. 13 (5), 467-472.

Ymous, A. et al. 2020. "I am just terrified of my future"—Epistemic Violence in Disability Related Technology Research. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-16.

7. Surveillance of Activist Movements

TOWARD SECURE SOCIAL NETWORKS FOR ACTIVISTS

Leah Rosenbloom

The Workshop School (USA)

leah.rosenbloom@workshopschool.org

ABSTRACT

Activists are a vulnerable population who risk personal harm to advocate for social, political, or environmental change. Recent technologies simultaneously threaten activists and facilitate their work. In particular, both activists and their antagonists use social media platforms to track movements for change. While some activists combat surveillance using end-to-end encrypted text and email, these tools do not provide adequate infrastructure to form and maintain connections between large groups of people. This paper presents an analysis of the current role of social media in activism, and explores new ways in which platform designers might leverage advancements in cryptography to build safer, more effective networking tools for activists.

INTRODUCTION

Activists are people who work to effect political, social, or environmental change. They seek to challenge and disrupt existing power structures, and have therefore been the frequent targets of surveillance, censorship, prosecution, and state violence (Churchill and Vander Wall, 1990; Poell, 2014; Breuer et. al., 2015; Rihl, 2020; The Moscow Times, 2021). This targeting not only threatens activists' lives and livelihoods, but also creates a chilling effect on the right of the people to petition the government for change, a cornerstone of democratic freedom.

Activists rely on community engagement, collective identity, and trust to organize successful movements for change (Della Porta, 2012; Mundt et al., 2018). We note the important distinction between activists, who organize based on community will to challenge or threaten *power structures*, and hate groups, who organize around discrimination and violence to challenge or threaten *people*. This work foregrounds the needs of activists, and considers how technologists might build tools that maximize the effectiveness of activist organizations while minimizing the ability of states and hate groups to leverage the same tools against marginalized communities.

Over the past decade, activists have increasingly organized ideas, groups, and actions over the internet, and in particular on social media platforms such as Facebook, Twitter, Instagram, and YouTube (Howard et al., 2011; Valenzuela, 2013; Bohdanova, 2014; Lee et al., 2016; Johnson, 2017; Mundt et al., 2018). Social media has become a powerful tool for activists and state powers alike: the ubiquitous and public nature of the platforms simultaneously creates infrastructure for organizing people and ideas, and for streamlining the surveillance and manipulation of those people and ideas.

Activists are thus caught in an uncomfortable position: as powerful as social media can be in support of organizing, it can facilitate intimidation and abuse just as effectively. While tools like Signal and PGP can provide secure, end-to-end encrypted alternatives to SMS and email, they depend on established relationships and context. Social media, on the other hand, is a networking tool that allows activists to form new relationships and build context collaboratively. One might expect an inverse relationship to exist between networking capabilities, which afford the visibility and accessibility necessary for

movements to scale, and security measures, which generally make networks inaccessible to anyone who is not already a trusted participant. This paper argues that secure social network developers might use advancements in anonymous authentication and searchable encryption to build a tool with the best of both worlds: the power of social media to engage the public and build community, and the power of cryptography to authenticate legitimate members of the movement and thwart state surveillance.

SOCIAL MEDIA AS A TOOL TO EFFECT CHANGE

Scholars in various disciplines have studied the impact of social media on movements for change. For the purpose of drawing connections, this paper includes analyses of six diverse movements: the Arab Spring (Howard et al., 2011), mass demonstrations for policy change in Chile (Valenzuela, 2013), the Euromaidan Uprising in Ukraine (Bohdanova, 2014), the Hong Kong Umbrella Movement (Lee et al., 2016), the Dakota Access Pipeline Protests in Standing Rock, North Dakota (Johnson, 2017), and the international Black Lives Matter movement (Mundt et al., 2018). Despite the different motivations, objectives, and structures of these movements, they all leveraged social media to engage new participants and effect change with unprecedented efficiency.

Facilitation of Traditional Organizing: Speed, Scope, and Scale

Researchers are careful to distinguish social media as a *facilitator*, rather than a direct or independent cause, of activist movements (Valenzuela, 2013; Bohdanova, 2014; Lee, 2016; Mundt et al., 2018). Social media platforms and other digital organizing mechanisms must therefore be considered tools that people can build, manipulate, and improve. Specifically, activists use social media networking to increase the *speed* and *scope* of information dissemination and participation, as well as to facilitate the organization of essentials like funding, food and water, legal assistance, and medical brigades. This efficiency allows movements to quickly *scale*.

During the Arab Spring, activists used social media to spread what Howard et al. called “freedom memes” that served to raise awareness of regime corruption and brutality and to encourage participation in ground protests (2011, p. 3). Bohdanova also noted the “circular relationship” between social media use and ground protest in Ukraine, arguing that social media galvanized protest, which in turn increased the demand for more social networking (2014, p. 137). Furthermore, the most important weeks of critical ground protests in Tunisia and Egypt were preceded by exponential spikes in trending hashtags on Twitter (Howard et al., 2011, p. 12, 17). This supports the theory that social media use at least fueled, if not catalyzed, the protests leading up to the desired resignation of leadership in both of those countries.

Throughout the Arab Spring, activists used digital platforms to share real-time video evidence of activity on the ground, thereby creating new visual material for local activists and international media alike to analyze, share, and discuss (Howard et al., 2011, p. 22). The capacity of digital organizing platforms to quickly disseminate video evidence of state brutality, poverty, environmental exploitation, and other such causes of protest is one extremely powerful mechanism unique to digital organizing. The visual evidence of police brutality against activists and civilians circulated on social media also played a key role in international public support for the protests in Standing Rock (Johnson, 2017, p. 162) and the Black Lives Matter (BLM) movement (Mundt et al., 2018, p. 3). Social and other forms of independent media can ensure that the truth of events and activities on the ground are spread among supporters, allowing activists to bypass mainstream media narratives.

In addition to helping activists organize collective action, social media can also be used to help sustain it. During the Euromaidan Uprisings, Bohdanova describes Facebook groups that coordinated legal representation and medical brigades to provide protesters with a safety net of assistance, solidarity, and mutual support. Bohdanova argues that the crowdsourcing of resources over social media was "the most crucial for sustaining Euromaidan over a long period of time" (Bohdanova, 2014, p. 138). Similarly, BLM activists describe the expression of community and shared experience over social media as "key to maintaining pace and enthusiasm for the cause" (Mundt et al., 2018, p. 7). Social media platforms thus facilitate key support networks, providing activists with access to critical resources. This support, Mundt argues, has allowed BLM to "scale up" and build both internal and cross-movement coalitions (p. 9). The power of digital organizing platforms to help foster and maintain intersectional coalitions is understudied in the literature, and a good candidate for future work in this area.

The Critical Importance of Expression and Connection

Two empirical studies of citizens' use of social media conducted by Valenzuela in Chile (2011) and Lee et al. in Hong Kong (2016) found that the social media behaviors most strongly predictive of eventual activism were 1. sharing information and expressing opinions, and 2. connecting with activists directly. Interestingly, both Valenzuela and Lee et al. distinguish between *receiving* and *expressing* opinions and information. While the passive consumption of political information was not found to predict protest activity, *sharing* the information was (Valenzuela, 2011, p. 16; Lee, 2016, p. 463). Therefore, any tool for digital organizing might provide at minimum an ability for participants to share content and connect directly with activists.

Bridging Local and Global Communities

Social media has given people unprecedented access to the thoughts, struggles, and activities of communities all over the world. The movement for water protection in Standing Rock, North Dakota was perhaps the most stunning example of a highly-local-turned-global protest that received attention, solidarity, and millions in funding from people in 95 different countries (Johnson, 2017, p. 166). On the #NoDAPL global Day of Action, activists held over 300 events in all 50 U.S. states and dozens of cities worldwide (p. 166-167). During the height of the protests, one supporter claimed on Facebook that the local sheriff's office was using Facebook to surveil the thousands of ground protesters, and asked other supporters to use the "check-in" feature at Standing Rock to flood the Sheriff's Department with decoy suspects (p. 164). While this action was shown to be ineffective counter-surveillance, it drew over 1.5 million participants.

As the world's leading movement with a focus on intersecting oppressions, BLM is perhaps the best suited to leverage social media to organize around collective liberation. Not only has the national chapter of BLM used social media to publicly endorse and amplify other economic, racial, social, and environmental causes, but Mundt et al. found that BLM organizers at the local level highlighted the fundamental importance of connecting with activists working to support LGBTQ+ people, women, immigrants, indigenous people, and other non-black people of color, asserting that "the rising tide lifts all boats" (p. 9). BLM organizers highlighted the importance of social media in helping to achieve their goals of broad, long-term coalitions with other activist groups.

Tales of Caution and Resilience

There are several major caveats to the benefits of social media as an organizing tool discussed in the sections above. At the entry level, oppressed populations and rural populations do not have equal access the internet, mobile phones, and cellular data. Ever resilient, activists in several of the studies invented workarounds that bypassed a lack of technological resources among protesters: during nation-wide network outages in Egypt, activists continued to organize online using "satellite phones and dialup connections to Israel and Europe" (Howard et al., 2011, p. 16); water protectors in Standing Rock, a remote location with inconsistent cellular network coverage, established "Facebook Hill" as the closest location with enough signal to post updates from the protest camp (Johnson, 2017, p. 162); and Euromaidan activists erected physical "IT Tents" at protests that offered "free Internet access and computer equipment to protesters" (Bohdanova, 2014, p. 138). Regardless, it is important to recognize that no matter how hard activists try to work around the resource gap, technological methods of engagement pose constraints with respect to people who are physically or economically isolated from technology.

For those who are able to participate in social media-amplified activism, there are well-evidenced threats of surveillance, censorship, disinformation, harassment, prosecution, and state-violence. The next section is devoted to a discussion of these threats. While the focus of this paper is largely on state surveillance and abuse, we recognize the concern highlighted in Mundt et al.'s study regarding civilian hate groups, who also use social media to stalk and harass activists (p. 12). Given that the state has significantly more physical and financial power (and often greater incentive) to threaten activists on a systematic basis, we choose to focus on the state, and argue that any defense against state power will apply sufficiently to the relatively lesser power of civilians.

SOCIAL MEDIA AS AN ARM OF THE STATE

State powers have used evidence gathered on social media to murder and prosecute activists in retaliation for civil disobedience (Breuer et al., 2015; Rihl, 2020; The Moscow Times, 2021). Unlike other new forms of protest surveillance technology such as drones, stingrays and facial recognition, activists engage with social media voluntarily, and with intimate knowledge of the risks and harms associated with both surveillance and misinformation on social media platforms. This section presents an overview of the roots of the problems of state surveillance, censorship, and disinformation for activists on social media.

Surveillance

Social media platforms provide law enforcement access to users' personal data both directly and indirectly (Mateescu et al., 2015). Facebook and Google have designated "Law Enforcement Request Systems" that streamline requests for data acquisition and removal (Facebook, 2020; Google, 2020). The exact criteria and frequency of these requests is not available to the public. In the U.S., law enforcement does not have to obtain a warrant to access even password-protected content and private messages shared over closed networks, since social media platforms are defined as third-party entities and are therefore free to disclose user data under the third-party doctrine (Mund, 2017). While the U.S. Supreme Court recently challenged the third-party doctrine with respect to historical cell-site location tracking in *Carpenter v. United States* (Gee, 2020), there are still many court cases ahead to settle what a "reasonable" expectation of privacy means in the context of social media.

In addition to formal channels, law enforcement may access activists' profiles and posts by joining the platform and viewing publicly available content. Because activists need visibility in order to engage new community members and volunteers, their groups and posts are frequently public. While many activists are aware of and concerned about this visibility, and recognize that it makes them vulnerable, they continue to use social media for all of the reasons discussed in Section 2. We further discuss the trade-off of safety and visibility for activists in Section 4.3.

Censorship and Regulation

Repressive states often block and censor access to social media platforms (Poell, 2014). Even states that allow unrestricted access to social media platforms often pressure social media companies to censor or suppress certain political content (The Moscow Times, 2021), and there is an ongoing debate about the regulation of free speech on social media (Fleischman and Rosenbloom, 2020).

Historically, states have used media regulation "in the public interest" to serve a political purpose and maintain power (Samples & Matzko, 2020). In light of this precedent, civil rights organizations strongly advocate for free speech protections, even when they do not agree with the speakers. For example, in 1977 a Jewish lawyer at the American Civil Liberties Union defended the right of neo-Nazis to protest in a historically Jewish neighborhood because he knew this protection would have to extend equally to marginalized groups (Goldberger, 2020). Blanket censorship or strict regulatory measures on social media would likely have wide-ranging consequences for vulnerable populations, who rely on these platforms to expose suffering and corruption.

Terrorists and Extremists

Another way for states to stifle activism is to declare a particular activist group to be terrorists or extremists. States intentionally obfuscate and overload these terms with implicit connotations of fear, danger, and violence. Because counter-terrorism operations are often highly-classified, states are able to wield this designation against activists without public-facing explanation. States then use the implied *connotations* of these labels to justify censorship, surveillance, and violence. The term 'extremist' is especially confounded, as by definition it simply means any belief or action that is not mainstream. In practice, state powers often use 'extremist' to refer to any belief or action that does not align with their own political or financial interests.

Recent examples of this playbook include the United Kingdom's classification of the climate change advocacy group Extinction Rebellion as an "extremist ideology" (Dodd & Grierson, 2020) and Russia's addition of opposition leader Alexei Navalny's political network to the list of "terrorist and extremist organizations" (French Press Agency, 2021). Despite the fact that these movements are non-violent, they are listed alongside neo-Nazis, Daesh, and al-Qaida. The ruling party in Russia is currently using the designation to prosecute and imprison political opponents, one of whom was given a two year prison sentence over state-critical Tweets (The Moscow Times, 2021).

Disinformation Campaigns

Transnational disinformation campaigns became an internationally-recognized issue in the months preceding the 2016 U.S. Presidential election, when Russian state actors used targeted advertising, bots, and local groups on social media to spread election misinformation and amplify racist and anti-immigrant rhetoric (Faris et. al., 2017). Like social media-enabled activist movements, social media-enabled disinformation campaigns are also unprecedented in speed, scope, and scale.

The threat of misinformation and state, partisan, and corporate actors posing as average users on social media has severely degraded activists' ability to engage legitimate supporters from a place of trust and community, which are both essential to the success of social movements (Della Porta, 2012). Along with surveillance and censorship, technologists must therefore consider *trust* as an essential feature to create and maintain on digital organizing platforms.

Corporate-State Power

The algorithms employed by social media companies, which are designed to propagate clickable (and therefore monetizable) content, have also been shown to suppress content related to activist movements in favor of more conservative viewpoints (Roose, 2020). It is also in social media companies' financial interest to comply with law enforcement requests and avoid incurring the wrath of powerful political figures who might punish them with fines, lawsuits, profit-limiting regulations, or outright national bans. The last several years have seen state powers leverage protected access to social media platforms to push their political agendas, sidestep community standards, silence dissent, and incite violence against their political opponents (Marantz, 2020; Frenkel, 2021; The Moscow Times, 2021). The result of the symbiotic relationship between state powers and "big tech" is a propaganda-industrial complex (Fleischman & Rosenbloom, in press), where state powers and social media companies work in tandem to profit from censorship, surveillance, and disinformation.

THE BALANCE OF POWER WITHIN ACTIVIST MOVEMENTS

Activist movements are as diverse as the people who comprise them. There are, however, key structures in activist movements that shape how people organize thoughts, actions, and decision-making. Underlying all of these structures is the power derived from shared experience and community trust. Activists must evaluate the structures and needs of the movement in order to strike a balance between individual safety and collective visibility.

Formal Structures and Hierarchies

Activism is a broad umbrella term to describe movement towards social, political, or environmental change. Activist scholar Micah White defines activism using four theories of change: change via collective human action (such as protesting), change via external material forces (such as stock market fluctuations), change via individual spiritual practice (such as mindfulness meditation), and change via cosmic forces (such the COVID-19 pandemic). Activists may choose to focus on one or many of these theories, depending on their objectives and environment. Objectives of activist movements, White argues, can be separated into three layers: influence, whereby activists aim to change the *behavior* of people in power (policy change); reform, whereby activists aim to change who *has* power (change in representation); and revolution, whereby activists aim to change *how* power functions (change in the fundamental structure of governance). (White 2020)

Movements themselves are made up of activists who operate at both individual and structural levels. Two basic organizational structures of social movements are vertical structures, where activists follow a strict hierarchy of power and decision-making, and horizontal structures, where power and decision-making are distributed equally among all participants (Piven, 2013). Both structures have benefits and drawbacks. Vertical organization allows for swift decision-making and mobilization, at the expense of dampening individual voices and creating a small handful of high-value targets. Horizontal organization models the democratic freedom that activists strive to create, but can also lead to convoluted or

inconsistent messaging and demands. The ideal dynamic between vertical and horizontal structure again depends on individual movements' objectives and environment.

Community Protection and Trust Infrastructure

Trust is a precursor to the success of democratic movements (Della Porta, 2012). Activists build trust using the techniques described in Section 2: sharing ideas and opinions, building community, and supporting participants with material resources like water, food, legal representation, and medical aid. Trust is especially important in horizontally-aligned organizations, where there is no designated leader to affirm goals and resolve conflicts. In light of surveillance and disinformation, activists will need a way to build and maintain flexible and resilient trust networks. How to create and maintain trust infrastructure among activists online, especially as movements scale up and participants are less likely to know one another offline, is an open question considered further in Section 5.3.

Safety v. Visibility

In the article "Be Safe or Be Seen?", Tetyana Lokot notes the delicate balance between safety and visibility for Russian opposition figures (2018). While at first security and visibility seem mutually exclusive, Lokot argues that Russian activists strike a balance with conspicuous, open-source security practices to keep sensitive communications safe, and strategic visibility to create a sense of transparency and community within the movement.

Lokot describes the challenge facing Russian opposition activists and their organization, the Anti-Corruption Foundation (FBK), as "normalised and internalised...networked authoritarianism" (p. 333-334). Under these conditions, FBK activists not only utilize anti-surveillance tools like Tor, VPNs, proxies, end-to-end encryption, and two-factor authentication for themselves, but they also create and post guides, video tutorials, and information bulletins online to help supporters learn how and when to adopt these tools as well (p. 339-340). The open and enthusiastic nature of FBK activists' security and digital literacy practices has fostered collaboration between activists and technology experts in Russia, leading to the development of new anti-surveillance tools (p. 341).

FBK activists further integrate safety and visibility with a policy Lokot calls radical transparency, by which activists regularly release financial reports, campaign and investigation strategies, and videos of activities online. The activists' unfettered openness serves to both encourage public trust in the movement and thwart the state's ability to wield surveillance as a weapon against them. In the offensive realm of radical transparency, FBK activists also conduct counter-surveillance, turning drone footage of financial corruption and police brutality into YouTube videos garnering millions of views (p. 342-343). Like organizational structure, the ideal balance of safety and visibility for activists depends on the movement's objectives and surrounding environment. For the political opposition in Russia, making a commitment to digital literacy and radical transparency is one way to set the opposition apart from existing norms of tight information control, secrecy, and corruption.

THE ROLE OF CRYPTOGRAPHY

In 2015, cryptographer Phillip Rogaway began his paper "The Moral Character of Cryptographic Work" with three revolutionary words: "Cryptography rearranges power." This statement is consistent with his proposed framing of mass surveillance as "an instrument of power...an apparatus of control" (p. 26) for which cryptography is the best countermeasure we know. Rogaway further implores

cryptographers to consider the political and ethical implications of new cryptographic protocols. In the spirit of his work, this section considers cryptography in the context of activism.

The Cryptographic Poles of Safety and Visibility

In the safety v. visibility paradigm, end-to-end (E2E) encryption is all the way on the side of safety. When used correctly (with safe key storage and where both parties agree not to share the decrypted communications with anyone else), E2E encryption ensures that the content of participants' communication is perfectly concealed from third parties. E2E encryption can therefore circumvent the problem of the third-party doctrine and reestablish a reasonable expectation of privacy on third-party platforms. The total lack of external visibility, however, would also prevent activists from doing community outreach and engagement online. Another drawback of E2E encryption is that while it conceals the *contents* of a message, it does nothing to conceal the *metadata*—the location, timing, and other identifiable information about the communicating parties.

At the visibility end of the paradigm is communication over unencrypted, unauthenticated channels. For example, Umbrella Movement activists in Hong Kong used the peer-to-peer mobile application FireChat to work around network outages. This app and others like it were found to be functional, but easy to surveil (Dalek et al., 2014). Social media platforms are similarly visible and easy to surveil. In addition, social media platforms provide no comprehensive means for authenticating users as belonging to a particular community, allowing states and hate groups to pose as legitimate members of activist groups.

Anonymous Networks

Anonymous networks utilize onion routing (Camenisch & Lysyanskaya, 2005) or cryptographic shuffling (Corrigan-Gibbs & Ford, 2010) to hide the personally-identifiable metadata of users' communications. Activists might use anonymous networks to conceal the metadata of the E2E encrypted communications discussed in Section 5.1, or to send anonymous communication blasts to the public. At the community outreach level, however, activists would not be able to build or maintain trust using anonymous networks. By design, activists' posts over anonymous networks would be indistinguishable from those of their antagonists. This would create confusion over mission statements and mobilizing instructions: supporters would not be able to distinguish legitimate calls to action from those created to mislead, confuse, or trap them in a dangerous situation.

Using Anonymous Authentication to Create Trust Guarantees

Anonymous authentication is a process by which activists might prove to one another that they are legitimate members of an organization without revealing any other identifiable information about themselves. One way to achieve anonymous authentication is to apply anonymous credentials (Camenisch & Lysyanskaya, 2004) over top of an anonymous network. Anonymous credentials systems function in two steps: a setup phase, in which activists would obtain credentials that authenticate them as belonging to a particular organization, and a use phase, in which activists would present the credential to others, proving their legitimacy within the organization without revealing anything else about themselves.

Where and how activists might obtain their credentials would depend on the needs of their particular movement. A more vertical organization might use *delegatable* anonymous credentials (Crites & Lysyanskaya, 2019), by which an authority figure (such as a national organizer) could verify and pass

along credentialing power to the next level down (such as a local organizer), and so on. A more horizontal organization might employ *decentralized* anonymous credentials (Garman et al., 2014), where credential-granting authority is distributed among all members of the group.

The precise information activists would need to present in order to prove their legitimacy during the setup phase would similarly depend on activists' desired balance of safety and visibility. An organization with a focus on community might offer credentials only to those who request access in person. An organization looking for added visibility might permit anyone to join, but restrict post privileges to those with a special type of credential. Credentials can also be revoked in the event that a group member violates community agreements. This revocation property would empower activists to establish and uphold their own standards for moderation in digital spaces.

Anonymous credentials are a promising cryptographic primitive for activist movements in that they would help activists reclaim power over their identities and communities online. While anonymous credentials are not currently used in the everyday setting, newer constructions have made great improvements in efficiency, and functional implementations are on the horizon.

Using Searchable Encryption to Organize Private Information

Anonymous credentials can help maintain the balance of safety and visibility for activists' *identities*, but they do not account for all of activists' organizational needs. While E2E encryption is efficient for sharing content over private and group messaging, it does not offer a means to efficiently *organize* and *edit* that shared content between large groups of people. Having a secure way to collaborate on content such as event plans, protest routes, presentations, guides, and videos would further empower activists to control how and when they reveal information to the public.

One way to maintain the organizational structure of information while hiding the contents of the files is using searchable symmetric encryption (SSE) (Bellare et al., 2007). This primitive would allow activists to store information tagged with encrypted keywords on an untrusted database server, then use the same encrypted keywords to search efficiently over the stored information. The subset of SSE constructions that would permit activists to add, delete, and update files from the database is called *dynamic* searchable encryption (Cash et al., 2014). For even more flexibility, activists might employ nearest neighbor queries (Elmehdwi et al., 2014), which would allow supporters to search for content, events, or activities that are closely related to search criteria such as location or event type.

Unlike anonymous credentials, however, all searchable encryption schemes come with some (variably bounded) *leakage* that is still the subject of ongoing research and debate. Cryptanalysts have exploited the leakage of searchable encryption schemes instantiated on real-world data to partially or in some cases totally reconstruct encrypted databases (Cash et al., 2015; Kornaropolous et al., 2019; Liu et al., 2014). It is unclear how these attacks, which depend on the frequency and co-occurrence of popular or heavily-queried keywords within the database, would apply to activists' data. Other encrypted data access protocols, such as private information retrieval (PIR) (Chor et al., 1995) and oblivious RAM (ORAM) (Goldreich & Ostrovsky, 1996) do not leak information about the queries or files. However, these protocols are far less efficient and do not necessarily provide leakage-free search functionality (Naveed, 2015). PIR or ORAM might be useful for a library of extra-sensitive documents that activists could retrieve directly by index rather than keyword search.

CONCLUSION

A recent study of 53,000 people from 53 countries found that 48% of respondents think "the power of big tech companies" is a major threat to democracy (Wintour, 2021). The right of people to come together and demand social, political, and environmental change—in other words, activism—is at the core of democratic practice. Currently, activists are forced to engage with "big tech" and especially social media platforms in order to organize effectively. Technologists who place value in democracy must therefore turn their attention to the needs and challenges of activism in the digital age. As Rogaway suggests, cryptography is one effective tool with the ability to shift power away from states and corporations, and place it back into the hands of the people.

KEYWORDS: activism, social media, surveillance, cryptography, inclusive privacy and security.

REFERENCES

Bellare, M., Boldyreva, A., & O'Neill, A. (2007). Deterministic and efficiently searchable encryption. In *Annual International Cryptology Conference* (pp. 535-552). Springer, Berlin, Heidelberg.

Bohdanova, T. (2014). Unexpected revolution: the role of social media in Ukraine's Maidan uprising. *European View*, vol. 13, pp. 133-142.

Breuer, A., Landman, T., & Farquhar, D. (2015). Social media and protest mobilization: Evidence from the Tunisian revolution. *Democratization*, *22*(4), 764-792.

Camenisch, J., & Lysyanskaya, A. (2004, August). Signature schemes and anonymous credentials from bilinear maps. In *Annual International Cryptology Conference* (pp. 56-72). Springer, Berlin, Heidelberg.

Camenisch, J., & Lysyanskaya, A. (2005, August). A formal treatment of onion routing. In *Annual International Cryptology Conference* (pp. 169-187). Springer, Berlin, Heidelberg.

Cash, D., Grubbs, P., Perry, J., & Ristenpart, T. (2015). Leakage-abuse attacks against searchable encryption. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security* (pp. 668-679).

Cash, D., Jaeger, J., Jarecki, S., Jutla, C. S., Krawczyk, H., Rosu, M. C., & Steiner, M. (2014, February). Dynamic searchable encryption in very-large databases: data structures and implementation. In *NDSS* (Vol. 14, pp. 23-26).

Chor, B., Goldreich, O., Kushilevitz, E., & Sudan, M. (1995, October). Private information retrieval. In *Proceedings of IEEE 36th Annual Foundations of Computer Science* (pp. 41-50). IEEE.

Churchill, W., & Vander Wall, J. (1990). The COINTELPRO papers. *Boston: South End*. Retrieved from http://chinhnghia.com/Cointelpro_Papers.pdf

Crites, E. C., & Lysyanskaya, A. (2019, March). Delegatable anonymous credentials from mercurial signatures. In *Cryptographers' Track at the RSA Conference* (pp. 535-555). Springer, Cham.

Dalek, J., Winter, P., Dranka, A., Crete-Nishihata, M., & Senft, A. (2014). Asia Chats: Update on Line, KakaoTalk, and FireChat in China. *The Citizen Lab.* Retrieved from https://citizenlab.ca/2014/07/asia-chats-update-line-kakaotalk-firechat-china/

Della Porta, D. (2012). Critical trust: Social movements and democracy in times of crisis. *Cambio. Rivista sulle Trasformazioni Sociali*, *2*(4), 33-43.

Elmehdwi, Y., Samanthula, B. K., & Jiang, W. (2014, March). Secure k-nearest neighbor query over encrypted data in outsourced environments. In *2014 IEEE 30th International Conference on Data Engineering* (pp. 664-675). IEEE.

Facebook (2020). Law Enforcement Online Requests. Retrieved from https://www.facebook.com/records/login/

Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E., & Benkler, Y. (2017). Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election. *Berkman Klein Center Research Publication*, *6*.

Fleischman, W., & Rosenbloom, L. (2020). Problems with Problematic Speech on Social Media. *ETHICOMP 2020*, 116. Retrieved from https://dialnet.unirioja.es/descarga/libro/769585.pdf#page=181

Fleischman, W., & Rosenbloom, L. (in press). Social Media and the Rise of the Propaganda-Industrial Complex. *ETHICOMP 2021*, 79.

French Press Agency (2021). Russia designates Navalny network as 'extremist organization'. *Daily Sabah*. Retrieved from https://www.dailysabah.com/world/europe/russia-designates-navalny-network-as-extremist-organization

Frenkel, S. (2021). The storming of Capitol Hill was organized on social media. *The New York Times*. Retrieved from https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html

Garman, C., Green, M., & Miers, I. (2014, February). Decentralized Anonymous Credentials. In *NDSS*.

Gee, H. (2020). Last Call for the Third-Party Doctrine in the Digital Age after Carpenter? *BUJ Sci. & Tech. L.*, *26*, 286.

Goldberger, D. (2020). The Skokie Case: How I Came To Represent The Free Speech Rights of Nazis. *American Civil Liberties Union (ACLU)*. Retrieved from https://www.aclu.org/issues/free-speech/rights-protesters/skokie-case-how-i-came-represent-free-speech-rights-nazis

Goldreich, O., & Ostrovsky, R. (1996). Software protection and simulation on oblivious RAMs. *Journal of the ACM (JACM)*, *43*(3), 431-473.

Google (2020). Law Enforcement Request System. Retrieved from https://lers.google.com/signup_v2/landing

Howard, P., Duffy, A., Freelon, D., Hussain, M., Mari, W. & Mazaid, M. (2011). Opening closed regimes. *Project on Information Technology and Political Islam*. Working paper 2011.1.

Human Rights Watch (2018, 26 February). China: Big Data fuels crackdown in minority region. Retrieved from https://www.hrw.org/news/2018/02/26/china-big-data-fuels-crackdown-minority-region#

Johnson, H. (2017). # NoDAPL: Social media, empowerment, and civic participation at standing rock. *Library Trends*, *66*(2), 155-175.

Kornaropoulos, E. M., Papamanthou, C., & Tamassia, R. (2019, May). Data recovery on encrypted databases with k-nearest neighbor query leakage. In *2019 IEEE Symposium on Security and Privacy (SP)* (pp. 1033-1050). IEEE.

Lee, F. Chen, H-T. & Chan, M. (2016). Social media use and students' participation in a large-scale protest campaign: the case of Hong Kong's Umbrella Movement. *Telematics and Informatics*, vol. 34, pp. 457-469.

Liu, C., Zhu, L., Wang, M., & Tan, Y. A. (2014). Search pattern leakage in searchable encryption: Attacks and new construction. *Information Sciences*, *265*, 176-188.

Lokot, T. (2018). Be Safe or Be Seen? How Russian Activists Negotiate Visibility and Security in Online Resistance Practices. *Surveillance & Society*, *16*(3), 332-346.

Mateescu, A., Brunton, D., Rosenblat, A., Patton, D., Gold, Z., & Boyd, D. (2015). Social media surveillance and law enforcement. *Data Civ Rights*, *27*, 2015-2027.

Marantz, A. (2020). Why Facebook can't fix itself. *The New York Times*, October 12, 2020. Retrieved from https://www.newyorker.com/magazine/2020/10/19/why-facebook-cant-fix-itself

Mund, B. (2017). Social media searches and the reasonable expectation of privacy. *Yale JL & Tech.*, *19*, 238.

Mundt, M., Ross, K., & Burnett, C. M. (2018). Scaling social movements through social media: The case of Black Lives Matter. *Social Media+ Society*, *4*(4), 2056305118807911.

Oliver, P., Marwell, G., & Teixeira, R. (1985). A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action. *American journal of Sociology*, *91*(3), 522-556.

Piven, F. F. (2013). On the organizational question. *The Sociological Quarterly*, *54*(2), 191-193.

Poell, T. (2014). Social media activism and state censorship. *Social media, politics and the state: Protests, revolutions, riots, crime and policing in an age of Facebook, Twitter and YouTube*, 189-206.

Rihl, J. (2020). 'If your mom can go in and see it, so can the cops': How law enforcement is using social media to identify protesters in Pittsburgh. *Public Source*. Retrieved from https://www.publicsource.org/pittsburgh-police-arrest-blm-protesters-social-media-facial-recognition/

Rogaway, P. (2015). The Moral Character of Cryptographic Work. *IACR Cryptol. ePrint Arch.*, *2015*, 1162.

Roose, K. (2020). Social Media Giants Support Racial Justice. Their Products Undermine It. *The New York Times*. Retrieved from https://www.nytimes.com/2020/06/19/technology/facebook-youtube-twitter-black-lives-matter.html

Samples, J. & Matzko, P. (2020). Social Media Regulation in the Public Interest: Some Lessons from History. *Knight First Amendment Institute at Columbia University*. Retrieved from https://knightcolumbia.org/content/social-media-regulation-in-the-public-interest-some-lessons-from-history

The Moscow Times (2021, January). Social Media Platforms Delete Russian Posts Promoting Navalny Protests – State Censor. *The Moscow Times*. Retrieved from https://www.themoscowtimes.com/2021/01/22/social-media-platforms-delete-russian-posts-promoting-navalny-protests-state-censor-a72701

The Moscow Times (2021). Navalny Ally Jailed 2 Years for Anti-Government Tweets – Mediazona. *The Moscow Times*. Retrieved from https://www.themoscowtimes.com/2021/04/16/navalny-ally-jailed-2-years-for-tweets-inciting-extremism-mediazona-a73626

Valenzuela, S. (2013). Unpacking the use of social media for protest behavior: The roles of information, opinion expression, and activism. *American behavioral scientist*, *57*(7), 920-942.

White, M. (2020). What every activist must know. Micah White speaks with Moral Voices at Tufts. *YouTube*. Retrieved from https://www.youtube.com/watch?v=-qGdVajGUxQ&ab_channel=MicahWhite

Wintour, P. (2021). US seen as bigger threat to democracy than Russia or China, global poll finds. *The Guardian*. Retrieved from https://www.theguardian.com/world/2021/may/05/us-threat-democracy-russia-china-global-poll

SOCIAL MEDIA AND THE RISE OF THE PROPAGANDA-INDUSTRIAL COMPLEX

Leah Rosenbloom, William Fleischman

The Workshop School (USA), Villanova University (USA)

leah.rosenbloom@workshopschool.org; william.fleischman@villanova.edu

ABSTRACT

The riots at the U.S. Capitol have thrown the impact of social media-spread propaganda into sharp relief. While the 2020 Presidential election was independently confirmed by thousands of state and local officials, part of the U.S. population still believes that the election was stolen. On January 6th, 2021, with help from white supremacist organizations and a handful of sympathetic Capitol police, hundreds of them carried out an armed insurrection against the U.S. government.

Proposals for the mitigation of social media-fueled extremism often include surveillance and censorship. These solutions not only harm legitimate movements for social change and democracy, but are also ineffective against terrorism and miss the root cause of the Capitol violence. Advertising-based technology services like social media are designed to profit from the wildfire-like spread of content with high shock value. They intentionally and systematically propagate anything clickable and therefore monetizable, creating wells of misleading information.

Moreover, the problem with white supremacist propaganda extends far beyond social media. Police participation in the riots as well as centuries of selective law enforcement and police brutality illustrate that institutional powers are not held accountable for criminal activity including murder, the incitement of violence and insurrection, subornation of treason, hate crime, fraud, and abuse of power for financial gain. Institutional lawlessness is often coupled with propaganda which serves to deflect attention from criminal activity and vilify political opposition. The financial incentives of social media companies and the political incentives of institutional powers have coalesced to create a new challenge to democracy: the propaganda-industrial complex. We argue that the proper response to the propaganda-industrial complex is not to censor, but rather to strengthen the foundations of democracy: to demand legally-binding transparency from technology companies, government, and law enforcement agencies, to protect and elevate the right to demonstrate against institutional lawlessness, and to create digital space for people to express and engage with ideas free from commodification and algorithmic bias.

INTRODUCTION

On January 6th, 2021, a violent mob breached the U. S. Capitol. The mob had been mustered that morning by the sitting U.S. President and, inflamed by his words and those of his enablers, had attempted to prevent by force the formal certification of his loss in the 2020 Presidential election. This violent act of insurrection, in which five people were killed and members of Congress were besieged and under threat, has once again stimulated intense debates over anti-terrorism policy, domestic surveillance, bias in law enforcement, and the role of social media in organizing and inciting violence.

This is a pivotal moment to reevaluate the regulation of social media, the investigation of domestic terror, and societal perception of democratic free speech and assembly. Lawmakers and technologists alike must draw the critical distinction between groups coordinating to commit violence and those exercising their right to speak freely for the purpose of social and political change.

Legitimate activist groups are harmed by pervasive surveillance and censorship. Civil liberties groups have long noted that "federal law enforcement already has powerful tools to investigate and prosecute acts of domestic terrorism without any new laws, and that importing the anti-terrorism framework risks creating broad and vague powers that could be used to go after activists or religious minorities" (Emons, 2021). The history of the NSA's PRISM program and the USA Patriot Act illustrates the propensity for "mission creep" in regard to electronic surveillance.

Furthermore, these measures are not effective tools against terrorism. The success of "needle-in-a-haystack" surveillance programs has been oversold. Without human intelligence to direct the inquiry, dragnet programs generate large numbers of false positives (Kirschner, 2015). During Congressional hearings after the 2013 Snowden revelations, NSA Director General Alexander said that U.S. surveillance programs helped prevent 54 cases of terrorism since 9/11. Closer evaluation found that many of these claims were exaggerated and, in one of the more serious and high-profile cases, a terrorist plot was thwarted as a result of "standard investigative techniques" rather than NSA surveillance (Landau, 2013). Meanwhile, the NSA collected data on millions of Americans.

Historically, surveillance programs have disproportionately targeted immigrants, activists, and people of color (Churchill and Vander Wall, 1990). During the Capitol riots orchestrated by white supremacists, federal law enforcement did not interfere. Rather than being infiltrated by law enforcement investigators, white supremacist groups are themselves host to law enforcement and military personnel who join out of sympathy for their aims (Westervelt, 2021).

This paper examines the relationship between industrial complexes and white supremacist ideology. We introduce the idea of the propaganda-industrial complex, comparable to the surveillance-industrial complex, that is created by the commodification of white supremacist messaging. We argue that the proper response to this combination of profit and state control is neither censorship nor surveillance, but rather the regulation of social media and imposition of restrictions on both corporations and law enforcement. In order to establish effective, sustainable solutions to industrial complexes, however, it is not enough to regulate technology, or even put in place more robust measures to deter and respond to domestic terrorism; we must also address systems of oppression that allow state powers and technology industries to profit from the perpetuation of white supremacist idelogy.

DEFINING THE PROPAGANDA-INDUSTRIAL COMPLEX

An industrial complex is a symbiotic union between a profit-based institution and social or political institution. The term was first coined in 1961 by U.S. President Dwight Eisenhower, who warned against the integration of U.S. military with arms dealers to create a military-industrial complex (NPR Staff, 2011). Researchers have studied the proliferation of many subsequent industrial complexes, among them the prison-industrial complex (Schlosser, 1998) and school-to-prison pipeline educational reform-industrial complex (Fasching-Varner et. al., 2014), the medical-industrial complex (Relman, 1980), and the surveillance-industrial complex (Ball & Snider, 2013). Though industrial complexes span a broad range of disciplines, they all include three key players: a power-hungry authority, a vulnerable population, and an industry that profits from them both.

The Attention Economy

In *The Attention Merchants*, Tim Wu describes the evolution of mass and systematic commodification of human attention, from the primitive beginnings of the advertising industry to the media capture finesse of "The Attention Merchant Turned President" (Wu, 2016). In particular, the rise of the clickbait

economy and "free" advertising-based services has turned the attention economy into a self-sustaining industry. Content with a high shock value is more likely to retain attention, desensitize users to the truth, and create a fertile environment for propaganda, shock doctrine capitalism, and kleptocracy (Klein, 2007).

Exploitation of the Attention Economy for Political Power

Social media's global-scale, black-box attention economy is uniquely positioned to serve powerful state institutions and political influencers. In 2016, state actors in Russia purchased targeted political advertisements on Facebook, spreading election misinformation and race-baiting propaganda in the U.S. (Calabresi, 2017). Conversely, after detaining opposition leader Alexei Navalny, the Russian government censored calls for protest on TikTok, YouTube, VKontake, and Instagram, claiming they would "illegally incite minors to attend unauthorized rallies" (The Moscow Times, 2021). State actors can thus use their power to exploit the attention economy or shut it down, influencing public perception and socio-political outcomes with little or no accountability.

Big Tech Profit + Dissemination of Propaganda = Propaganda-Industrial Complex

Companies specializing in cameras, spyware, and artificial intelligence provide state powers with new infrastructure to bolster and expand state surveillance capabilities. Similarly, ubiquitous social media platforms provide state powers with new infrastructure to capture attention and spread propaganda. While the former involves an on-the-books trade, tax dollars for spyware, the latter is a more insidious form of capitalism that allows the companies to monetize the propaganda passively via third-part advertisers, concealing ties to state officials and further convoluting the regulatory process. In addition to using the platforms to spread propaganda, state powers can restrict or manipulate the flow of information by threatening the company's bottom line, either directly by censoring content, or indirectly by publicly condoning or vilifying groups or individuals. Social media companies and state powers alike profit from the spread of propaganda, creating a propaganda-industrial complex.

A CALL FOR AN INTERSECTIONAL PERSPECTIVE

Government regulation of social media is long overdue. Technology professionals and public welfare advocates have called for the suppression of election misinformation (Eisenstat, 2020), incitement to violence (Fleischman and Rosenbloom, 2020), and baseless conspiracy theories (Pazzanese, 2020), as well as the designation of social media companies as content-responsible publishers rather than as neutral "internet intermediaries" that merely host third-party content (Eisenstat, 2021). While these regulatory fronts are important, they do not address two key features of the propaganda-industrial complex: the platforms' underlying algorithms and their ability to profit from propaganda.

All the Moderators in the World Could Not Stem the Flow of Clickbait

The machine-learning algorithms employed by social media companies are specifically designed to propel the most shocking, monetizable content to the forefront of targeted users' feeds. Assigning a team of human beings to the task of moderating the overwhelming volume of content on these platforms is not only traumatizing (Newton, 2019), but also futile. The proliferation of extreme content exists and is promoted by algorithmic design; moderators can ban one user to take down a hundred violent videos. The next most shocking content will float right to the top.

Content Regulation is Subject to Political Whims

In a comprehensive case study of media regulation in the U.S., Samples and Matzko argue that media regulation "in the public interest" is often used to serve a political purpose (2020). Both major U.S. political parties have complained about social media platform policies. Republicans contend they are being censored while Democrats decry platform policies that favor conservative media outlets that disseminate false and tendentious material (Lima et al., 2020). The ruling party in Russia recently used their power as regulators to demand that social media companies remove posts supporting opposition protests (The Moscow Times, 2021). Regulation of the many gray areas fraught with partisan political interest might therefore serve the propaganda-industrial complex rather than the public.

Regulation of Algorithms and Advertising is Better Than Regulation of People and Content

Given the limits of content moderation, it would be useful to focus our first regulatory efforts on the underlying algorithms and advertising models that purposefully amplify content with high shock value. Effective regulation of the machine-learning algorithms must start with two-fold transparency: companies must not only reveal the structure of the algorithm, but also an (anonymized) summary of the data they are collecting and using to train the algorithm. Regulatory bodies need this information to understand exactly how content is spreading and to whom.

Responding to criticism about its role in the 2016 U.S. presidential election, Facebook pledged to establish a publicly searchable archive of political ads. Advertisers are supposed to undergo various identity verification measures, and to put certain disclosures on their ads. The ad library is supposed to identify who is paying for each ad, how much the ad costs, and demographic information about users who see the ad. However, advertisers' microtargeting requests are not disclosed.

Studies of the archive have uncovered shortcomings. In several instances, ads sought by researchers disappeared from the archive (Kelly et al., 2019). One study indicated that "Facebook failed to identify and label 9.7% of ads for elections and issues placed between May 2018 and June 2019. These undisclosed ads represented $37 million in spending." And in the 2020 presidential election, nearly 10% of ads ran without an initial disclosure (Silverman & Mac, 2020).

In addition, Facebook and other social media have been sued by the Attorney General of the State of Washington for repeated violation of the Washington State campaign finance laws that require social media companies to maintain records about political ads on their platforms and make those records available for public inspection (Bishop, 2020).

One sensible step in regulating the activity of social media around elections and political campaigns would be to convert Facebook's gracious gesture into a rigorous legal requirement that any social media platform maintain an accurate archive of political ads including information as to the source (and measures taken to confirm the legitimacy of the source), cost, and any microtargeting specifications regarding each ad. A clear picture of how algorithms, data collection, and data monetization function is an essential precondition for regulators to formulate effective oversight measures that can address the problem at its root.

The Urgency of an Intersectional Perspective

The world is currently experiencing a rise in white nationalism (Becker, 2019). While it is true that social media companies perpetuate, amplify, and profit from these trends, it is also not reasonable to expect social media companies to eradicate them. The propaganda-industrial complex is a highly

effective tool for the spread of white supremacist ideology, but it is not the root cause; even if social media platforms were demonetized and presented differing points of view, institutional lawlessness and systemic oppression would persist. During the January 6th riots, some Capitol Police openly welcomed the rioters into the Capitol, removing barricades and taking selfies with them; some law enforcement officers and military officials from different places around the country traveled to participate in the riots, and many more showed solidarity and support on social media (Westervelt, 2021). It is therefore critical for society through regulation, law, and norms, to address the institutional and financial systems of oppression that create a permissive environment for white supremacist ideology—not only among citizens, but within the government and throughout law enforcement agencies.

Acknowledging the Difference Between Institutional Violence and the Movements Trying to End Institutional Violence

Black Lives Matter (BLM) is a transnational movement against violence driven by white supremacist ideology. Despite the fact that neither local nor national-level BLM organizers have ever advocated or condoned violence, conservative politicians and media outlets in the U.S. have attempted to equate them with the white supremacist organizers of the Capitol riots (Brantley-Jones, 2021). Moreover, while the BLM movement has popular support in the U.S., a Facebook-owned data company reported that seven out of ten widely-circulated posts about BLM on its platform were "critical of the movement" (Roose, 2020). Despite the challenges of disproportionate representation on both traditional and social media, BLM and other organizations advocating for social change have leveraged these platforms to share ideas, recruit volunteers, and organize collective action (Mundt et al., 2018). It is important that lawmakers and technologists alike respect the necessary work of groups like BLM to dismantle white supremacist systems of oppression, and do not let regulation or platform design impede these groups' right to free speech and assembly.

The past few years has seen a notable rise in activism among highly-skilled technical employees of the major social media platforms. At Google, this activism sometimes resulted in termination of employment for individuals who voiced displeasure about company policies, but has borne fruit in the recent organization of a new union at the company (Ghaffary, 2021). At Facebook, over 200 workers revolted after public disagreement with Mark Zuckerberg's decision not to take down a post by President Trump concerning protests around the murder of George Floyd. Scientists and civil rights organizations backed the employees in criticizing Facebook for its role in disseminating false information, culminating in the creation of the Stop the Hate for Profit movement. (Alon-Beck, 2020)

CONCLUSIONS

The recently announced decision of the quasi-independent Facebook Oversight Board (Facebook's so-called "Supreme Court") concerning the suspension of Donald Trump provides a revealing lens through which to reconsider the main aspects of the power imbalances at the heart of the propaganda-industrial complex. The board upheld the suspension of the former U.S. President's account, finding that his posts had "severely violated" Facebook's rules. However, it criticized the platform for imposing an "indeterminate and standardless penalty of indefinite suspension", stating Facebook should "publicly explain the rules that it uses when it imposes account-level sanctions against influential users." (Axios, 2021)

Major social media platforms have been criticized for promulgating rules for permissible speech that are troublingly vague and inconsistently applied (Fleischman and Rosenbloom, 2020). The Facebook

Oversight Board's finding in the current case is eerily reminiscent: "Facebook cannot make up the rules as it goes, and anyone concerned about its power should be concerned about allowing this. Having clear rules that apply to all users and [to] Facebook is essential for ensuring the company treats users fairly." (Axios, 2021)

The Facebook-Trump misadventure can be traced to a 2015 decision by Facebook executives not to take down a post of then-candidate Donald Trump that violated the platform's Community Standard prohibiting "all content that directly attacks people based on race, ethnicity, national origin, or religion." Trump had posted a call for "a total and complete shutdown of Muslims entering the United States" and insinuated that all (1.8 billion) Muslims "have no sense of reason or respect for human life." The exception allowing the post to stand was made in deference to Trump's prominence and to avoid incurring his wrath and that of his followers. In the words of one policy executive, "Don't poke the bear." (Marantz, 2020)

Apart from the grievous miscalculation that, given the precedent of acquiescence in this one instance, such transgressive speech could subsequently be restrained, the notion of giving greater leeway in standards of speech to prominent individuals (and, as matters turned out, to the leader of one of the world's most powerful nations) seems entirely wrong-headed. Such individuals have incomparably greater resources for broadcasting their opinions than ordinary citizens. In our view, platform speech policies should be uniformly enforced.

If there is a case for greater leeway, it ought perhaps to be extended to those in positions of inferior power. For a good example, we might turn to that model democracy, post-Soviet Russia. Here, social media companies have capitulated to Russian President Vladimir Putin, who has at his disposal the entire apparatus of state-run media, in taking down the posts of dissident Alexei Navalny and his supporters relating specifically to non-violent protests (The Moscow Times, 2021).

In the same way, clear rules evenly applied might help to put to rest the mutual fears of major political parties that social media platforms are doing them dirt in favor of their opponents. At any rate, with obvious election fraud and explicitly violent rhetoric ruled out, the parties will have to be more subtle and ingenious in exploiting the gray areas of political misrepresentation and misinformation. (With a single voice they respond: "We can do that!")

KEYWORDS: regulation of social media, activist movements, propaganda, surveillance.

REFERENCES

Alon-Beck, A. (2020). The Facebook Saga: When Tech Employees Revolt. *Forbes Magazine,* July 22, 2020. Retrieved from https://www.forbes.com/sites/anatalonbeck/2020/07/22/the-facebook-saga-when-tech-employees-revolt/?sh=71717ec16c32

Axios (2021). Oversight Board upholds Trump's Facebook suspension. *Axios*, May 5, 2021. Retrieved from https://www.axios.com/facebook-trump-ban-oversight-board-bc239ed4-3b42-4d94-960d-078da5deec3f.html

Ball, K., & Snider, L. (Eds.). (2013). *The surveillance-industrial complex: A political economy of surveillance*. Routledge.

Becker, P. (2019). White nationalism, born in the USA, is now a global terror threat. *The World. Retrieved from https://www.pri.org/stories/2019-03-20/white-nationalism-born-usa-now-global-terror-threat*

Bishop, T. (2020). Twitter to pay $100k to Washington state in settlement over political ad disclosure violations. *GeekWire,* October 13, 2020. Retrieved from https://www.geekwire.com/2020/twitter-agrees-pay-100k-washington-state-settlement-political-ad-disclosure-violations/

Brantley-Jones, K. (2021). False equivalency between Black Lives Matter and Capitol siege: Experts, advocates. *ABC News*. Retrieved from https://abcnews.go.com/US/false-equivalency-black-lives-matter-capitol-siege-experts/story?id=75251279

Calabresi, M. (2017). Inside Russia's Social Media War on America. *Time Magazine.* Retrieved from https://time.com/4783932/inside-russia-social-media-war-america/

Churchill, W., & Vander Wall, J. (1990). The COINTELPRO papers. *Boston: South End*. Retrieved from http://chinhnghia.com/Cointelpro_Papers.pdf

Eisenstat, Y. (2020). How to Combat Online Voter Suppression. *The Brookings Institution, TECH STREAM.* Retrieved from https://www.brookings.edu/techstream/how-to-combat-online-voter-suppression/

Eisenstat, Y. (2021). How to Hold Social Media Accountable for Undermining Democracy. *Harvard Business Review*. Retrieved from https://hbr.org/2021/01/how-to-hold-social-media-accountable-for-undermining-democracy

Emons, A. (2021). Capitol Hill Assault Revives Calls for Domestic Terrorism Law, But Civil Liberties Groups are Wary. *The Intercept,* January 10, 2021. Retrieved from https://theintercept.com/2021/01/10/capitol-hill-riot-domestic-terrorism-legislation/

Fasching-Varner, K. J., Mitchell, R. W., Martin, L. L., & Bennett-Haron, K. P. (2014). Beyond school-to-prison pipeline and toward an educational and penal realism. *Equity & Excellence in Education*, *47*(4), 410-429.

Fleischman, W., & Rosenbloom, L. (2020). Problems with Problematic Speech on Social Media. *ETHICOMP 2020*, 116. Retrieved from https://dialnet.unirioja.es/descarga/libro/769585.pdf#page=181

Ghaffary, S. (2021). Google's new union, briefly explained. *Vox: Recode,* January 4, 2021. Retrieved from https://www.vox.com/recode/22213494/google-union-alphabet-workers-tech-organizing-activism-labor

Kelly, J., Blood, D., & O Murchu, C. (2019). Facebook under fire as political ads vanish from archive, *Financial Times.* Retrieved from https://www.ft.com/content/e6fb805e-1b78-11ea-97df-cc63de1d73f4

Kirschner, L. (2015). What's the Evidence that Mass Surveillance Works? Not Much. *ProPublica.* Retrieved from https://www.propublica.org/article/whats-the-evidence-mass-surveillance-works-not-much

Klein, N. (2007). *The shock doctrine: The rise of disaster capitalism*. Macmillan.

Landau, S. (2013). Making Sense from Snowden. *IEEE Security and Privacy*, July/August 2013, pp. 66-75. Retrieved from https://privacyink.org/pdf/MakingSense.pdf

Lima, C., Overly, S., Niedzwiadek, N. & Nylen, L. 'Censorship teams' vs. 'working the refs': Key moments from today's hearing with tech CEOs. *Politico.* Retrieved from https://www.politico.com/news/2020/11/17/facebook-twitter-senate-tech-hearing-436975

Marantz, A. (2020). Why Facebook can't fix itself. *The New York Times*, October 12, 2020. Retrieved from https://www.newyorker.com/magazine/2020/10/19/why-facebook-cant-fix-itself

Mundt, M., Ross, K., & Burnett, C. M. (2018). Scaling social movements through social media: The case of Black Lives Matter. *Social Media+ Society*, *4*(4), 2056305118807911.

Newton, C. (2019). The Trauma Floor: The secret lives of Facebook moderators in America. *The Verge.* Retrieved from https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

NPR Staff (2011). Ike's Warning of Military Expansion, 50 Years Later. *National Public Radio (NPR), Morning Edition.* Retrieved from https://www.npr.org/2011/01/17/132942244/ikes-warning-of-military-expansion-50-years-later

Pazzanese, C. (2020). Battling the 'pandemic of misinformation'. *The Harvard Gazette, Health & Medicine.* Retrieved from https://news.harvard.edu/gazette/story/2020/05/social-media-used-to-spread-create-covid-19-falsehoods/

Relman, A. S. (1980). The new medical-industrial complex. *New England Journal of Medicine*, *303*(17), 963-970.

Roose, K. (2020). Social Media Giants Support Racial Justice. Their Products Undermine It. *The New York Times*. Retrieved from https://www.nytimes.com/2020/06/19/technology/facebook-youtube-twitter-black-lives-matter.html

Samples, J. & Matzko, P. (2020). Social Media Regulation in the Public Interest: Some Lessons from History. *Knight First Amendment Institute at Columbia University*. Retrieved from https://knightcolumbia.org/content/social-media-regulation-in-the-public-interest-some-lessons-from-history

Schlosser, E. (1998). The prison-industrial complex. *The Atlantic Monthly*, *282*(6), 51-77.

Silverman, C. & Mac, R. (2020). Facebook Promised To Label Political Ads, But Ads For Biden, The Daily Wire, And Interest Groups Are Slipping Through. *Buzzfeed News.* Retrieved from https://www.buzzfeednews.com/article/craigsilverman/facebook-biden-election-ads

The Moscow Times (2021, January). Social Media Platforms Delete Russian Posts Promoting Navalny Protests – State Censor. *The Moscow Times*. Retrieved from https://www.themoscowtimes.com/2021/01/22/social-media-platforms-delete-russian-posts-promoting-navalny-protests-state-censor-a72701

The Moscow Times (2021, April). Navalny Ally Jailed 2 Years for Anti-Government Tweets – Mediazona. *The Moscow Times*. Retrieved from https://www.themoscowtimes.com/2021/04/16/navalny-ally-jailed-2-years-for-tweets-inciting-extremism-mediazona-a73626

Westervelt, E. (2021). Off-Duty Police Officers Investigated, Charged With Participating In Capitol Riot. *National Public Radio (NPR)*. Retrieved from https://www.npr.org/2021/01/15/956896923/police-officers-across-nation-face-federal-charges-for-involvement-in-capitol-ri

Wu, T. (2017). *The attention merchants: The epic scramble to get inside our heads*. Vintage.

8. What Will Cybersecurity's “New Normal” Look Like?

USING A SECURITY PROTOCOL TO PROTECT AGAINST FALSE LINKS

Sabina Szymoniak

Czestochowa University of Technology (Poland)

sabina.szymoniak@icis.pcz.pl

ABSTRACT

The article discusses the problem of internet users protection against false URLs. During the coronavirus pandemic, we realized the value and benefits of remote work. Also, training began to be conducted remotely. Available webinars or training made it possible for us to broaden our knowledge. Organizers of such activities can send or share the proposals for participation via social networks or e-mail messages. The development of remote activities contributed to an increase in cybercrimes. A specially crafted script may be hidden under a seemingly safe URL for an online meeting. We propose a new security protocol for users authentication and protection against false links. The protocol consists of two parts: initial and verification. In the initial part, the user must agree on his unique identifier with the trusted Distribution Center. The event organizer must perform a similar action. In the verification part, the meeting participant will verify the correctness of the connection URL. This protocol implements AAA-logic. We examined the security of our protocol using a verification tool. This tool enables verification of security protocols, including various time parameters. The results obtained are promising. Further, we plan to develop the proposed protocol and prepare a system that will use it.

INTRODUCTION

The coronavirus pandemic has shown the value and benefits of remote work. During the binding restrictions, people had to find themselves in the new reality. They had to keep working and performing their official duties. Also, children and other learners moved to the virtual world to continue their education. As time passed, scientific conferences and training began to be conducted remotely too. Thanks to available webinars or training, we can broaden our knowledge. Organizers of such activities very often send and share proposals for participation via social networks or e-mail messages.

Unfortunately, with the development of remote activity, an increase in cybercrimes can also be observed. People who respond to an invitation to a webinar or training may also be exposed to criminals. A specially crafted script may be hidden under a seemingly safe URL for an online meeting. Running this script after clicking on a link can cause enormous damage. The damages include the installation of malware on our device or the hijacking of login data by cybercriminals.

Security is an essential element of everyone's life. There are two solutions to protect against such situations. The first is the detection of false links, using artificial intelligence methods. In this case, there are many methods to accomplish detection which URL contains deceptive content. Yang et al. (Yang et al., 2020) observed that vulnerabilities in recommender systems may encourage deliberate manipulation by malicious users. Also, Lai et al. (Lai et al., 2020) looked at the impact of social recommendation systems on the presence of malicious URLs on social networks. In (Song et al., 2020), the authors presented a novel method of detecting malicious code in JavaScript, based on deep learning. Baccouche et al. (Baccouche et al., 2020) have proposed an LSTM model that will identify

mischievous text regardless of the source. Also, in (Ispahany et al., 2020), (Xiao et al., 2020) and (Patel et al., 2020), we can find other methods to the detection of false links.

The second solution is the use of security protocols. These protocols are short algorithms that make it possible to achieve security goals. The two most meaningful goals are mutual user authentication or the exchange of confidential information. Over the past decades, many different security protocols have been developed. It is worth mentioning the Needham Schroeder protocol (Needham et al., 1978), which has been used for many years for user authentication.

With the advent of new security protocols, various methods of protocols verification and tools enabling them appeared. Worth mentioning are methods and tools presented in ((Dolev et al., 1983), (Burrows et al., 1989), (Lowe, 1996), (Paulson, 1999), (Armando A. et al., 2005), (Nigam et al., 2016), (Blanchet B., 2016), (Chadha et al., 2017), (Basin et al., 2018), (Siedlecka-Lamch et al., 2019)). Also, in the case of security protocols, it is worth mentioning the following papers (Steingartner et al., 2019), (Galinec et al., 2019), (Radaković et al., 2018), (Čibej et al., 2019),(Piątkowski, 2020).

In this article, we propose a new security protocol. We can use our protocol for users authentication and protection against false links. The protocol consists of two parts: preparatory and verification. In the preparatory part, the user must agree on his unique identifier with the trusted Distribution Center. Also, the event organizer must perform a similar action. In the verification part, the user (who wants to participate in any event) will verify the correctness of the URL used to make the connection.

Our protocol aims to implement AAA (Authentication, Authorization, Accounting) logic. This logic relates to a dedicated security framework used to mediate network and application access. Authentication describes how a user is identified to a network or application. Authorization is linked to the policy enforcement process for users. Accounting records session statistics and user information. Thanks to this solution, it is possible to control access to resources, including audit rules. Only authorized users will have access to the network, resources or applications (Galinec et al., 2019), (Steingartner et al., 2021).

The rest of the article is organized as follows. In the second Section, we present Related Works. The next Section presents our new security protocol. In the fourth Section, we present our experimental results. The last Section includes conclusions and plans for the future.

RELATED WORKS

One of the significant security protocols is Needham Schroeder Public key protocol (Needham et al., 1978). This protocol aims to mutual users authentication. In (Lowe, 1996), Gavin Lowe indicated a simple way to attack NSPK. Also, the worth mentioning about Wide Mouthed Frog protocol (Burrows et al., 1989). This protocol aims to the distribution of a new shared symmetric key. WMF is vulnerable to a replay attack.

Next to security protocol is MobInfoSec (Siedlecka-Lamch et al., 2019), (Siedlecka-Lamch, 2020). This protocol enables encryption and sharing of confidential information. It can be used by a group of connected users of mobile devices. Also, we use security protocols during voice calls, video calls, and chat instant messaging. Signal protocol is an example of this usage. This cryptographic protocol can be used to provide end-to-end encryption. Another application of security protocols is in Wireless Sensor Networks (Maheswari et al., 2021).

Also, security protocols can be used in IoT systems and Wireless Sensors Networks. In (Ko et al., 2021) authors proposed a security protocol for securing communication between Unmanned Aerial Vehicles. This protocol provides confidentiality and non-repudiation, which are essential for secure military

communication. In (Kwon et al., 2021) suggest a secure and lightweight mutual authentication protocol for Wireless Sensors Networks. This protocol provides mutual authentication and forward secrecy. In (Moreno-Cruz et al., 2020), the authors provide a new security protocol that complies with the Wireless Communication Protocol standard. This protocol takes into account the cross-layer principle to use power consumption to dynamically adapt its operation, which goes from a pure effort to near real-time.

A NEW SECURITY PROTOCOL

The protocol we propose used to establish unique user identifiers and authenticate them during virtual meetings. There are four users in it. The first one is the meeting participant. He is a user who has a link to the event and wants to participate in it. The other user is the meeting organizer, that is, the user who is preparing the event. The third participant is the Distribution Center. It is an application that is responsible for the distribution of unique user identifiers. The last participant in the protocol is the Authentication Center, which is also an application. This application is responsible for verifying participants in virtual events.

The main goals of the proposed protocol are:

- the reconciliation of unique user identifiers,
- the mutual authentication of users,
- the distribution of a symmetric key.

Figure 1. The first part of the proposed security protocol.

$$
\begin{aligned}
\alpha_1 \quad & A \rightarrow DC: \quad \{i(A), T_A\}_{K_A^-} \\
\alpha_2 \quad & DC \rightarrow AC: \quad \{\#hash(\{UID_A, i(A)\}_{K_A^+}), \\
& \qquad T_{DC}^A\}_{K_{DC-AC}} \\
\alpha_3 \quad & AC \rightarrow A: \quad \{\#hash(\{UID_A, i(A)\}_{K_A^+}), \\
& \qquad T_{DC}^A\}_{K_A^+} \\
\alpha_4 \quad & A \rightarrow AC: \quad \{T_{DC}^A\}_{K_{AC}^+}
\end{aligned}
$$

Source: self-elaboration

Our protocol consists of a preparatory and verification part. In the first part, users establish their unique user identifiers from the Distribution Center. During the second part of the protocol, users can verify the identity of the other user. A potential meeting participant can validate the meeting URL and, therefore, the identity of the organizer. On the other hand, the meeting organizer can verify the identity of a potential event participant. After correct verification, participants agree on a symmetric key that will be used for their further communication.

Figure 1 shows the syntax of the proposed protocol first part in the *Alice-Bob* notation. By *A*, we marked the user who wants to participate in the event or is its organizer. *DC* stands for Distribution

Center that generates and sends individual user identifiers. We marked the Authentication Center, which confirms the identity of the meeting participants, via *AC*.

The first part of this protocol consists of four steps. In the first step, the user sends a request to Distribution Center to generate an individual user identifier for him. This step should be performed by both the meeting organizer and the participant. The message includes a text user identifier and a timestamp generated by the user. For the meeting participant, the text identifier will contain his name and surname. For the meeting organizer, the text identifier will contain the URL of the event. The user must encrypt this message with his private key.

In the second step, the Distribution Center must generate a numeric, unique user identifier (UID_A) and a timestamp generated by Distribution Center that will certify its validity. Then, both numeric and text identifiers are encrypted with the meeting participant's public key. The resulting Distribution Center ciphertext is subjected to a hash function and then added to the message with the generated special timestamp. The resulting message is encrypted with a symmetric key shared between the Distribution Center and Authentication Center and sent to AC.

Figure 2. The second part of the proposed security protocol

$$
\begin{aligned}
\alpha_1 \quad & A \rightarrow AC: && \{\#hash(\{UID_A, i(A)\}_{K_A^+}), \\
& && T_{DC}^A, i(B)\}_{K_{AC}^+} \\
\alpha_2 \quad & AC \rightarrow A: && \{\{\#hash(\{UID_B, i(B)\}_{K_A^+}), T_{AC}^B\}_{K_A^+}, \\
& && \{\#hash(\{UID_A, i(A)\}_{K_B^+}), T_{AC}^A\}_{K_B^+}\}_{K_A^+} \\
\alpha_3 \quad & A \rightarrow B: && \{\#hash(\{UID_A, i(A)\}_{K_B^+}), T_{AC}^A\}_{K_B^+} \\
\alpha_4 \quad & B \rightarrow A: && \{T_{DC}^A, T_{DC}^B, K_{AB}\}_{K_A^+}
\end{aligned}
$$

Source: self-elaboration.

After decrypting the message, Authentication Center saves the meeting participant's data in its database. Each such entry will help AC in user authentication. Then the Authentication Center sends to A an identical message as in the previous step. Authentication Center encrypts this message with public key A (step 3).

After decrypting the messages from step three, A saves its data. Then, A constructs a message containing the timestamp generated by DC. A encrypts this message with the Authentication Center public key and forwards it to Authentication Center (fourth step). Thanks to this, A confirms its identity with AC and DC.

Figure 2 shows the syntax of the proposed protocol second part in *Alice-Bob* notation. Compared to the first part of the protocol, a new participant appears here, denoted as B. We can consider this part of the protocol in two ways, depending on the user's role in initiating the protocol. If A is a potential meeting participant, then B is the organizer. In that case, A checks the validity and validity of the URL. Conversely, A, as the meeting organizer, verifies the identity of the event participant.

In the first step of this part of the proposed protocol, A want to verify the identity of user B. He sends a request to the Authentication Center. This message contains A's unique identifier (hashed and encrypted with A public key) and a timestamp generated by Distribution Center for A and text identifier B. The text identifier may be either a URL address or the name and surname of the participant. This situation depends on the role of A in the protocol execution. A message prepared in this way encrypts the AC with the public key and sends it to Authentication Center.

Before the execution of the second step, the Authentication Center checks the identity of A. If such a user is in the AC database, it verifies the identity of B. After both correctly verified, AC constructs a message for A. This message consists of two ciphertexts. The first one contains B's unique identifier and its timestamp. This ciphertext is encrypted with the public key A. The second ciphertext contains A's unique identifier and his timestamp. This ciphertext is encrypted with the public key B. This means that the first part of the complex message is intended for A and only A can decrypt it, while the second part is intended for B, and only B can decrypt it. The entire message is encrypted with A's public key and sent to him.

In case of incorrect verification of one of the users, the Authentication Center sends ciphertexts filled with zeros. Thus, the user will know that the URL of the meeting is false, and the organizer will know that the user does not exist.

In the third step of the proposed protocol, A sends to B the second ciphertext of the earlier message. After reading it, B learns A's unique identifier and its timestamp. He may proceed to confirm his identity to B.

In the last step, B sends a message to A containing the timestamps of both users. Additionally, it generates a symmetric key and puts it in the message. The generated key will be used in further communication of these users. He encrypts this message with A's public key.

EXPERIMENTAL RESULTS

We then conducted a preliminary security study of our protocol. We used the model and tools described in (Szymoniak, 2010), (Szymoniak, 2021). Referring to the mentioned methodology, we conducted a series of tests related to the analysis of protocol execution times and time simulations of the protocol's operation in a computer network. The analyzes take into account the presence of an Intruder (Dolev et al., 1983). The intruder is a dishonest user who wants to intercept the confidential data of other users. We performed the protocol analysis using a computer with Linux Ubuntu operating system, Intel Core i5 processor and 16 GB RAM. We used an abstract time unit ([tu]) to determine the time.

According to the mentioned methodology of research, we used the following values:

- minimum (D_{min}), current (D) and maximum (D_{max}) delay in the network value,
- minimum (T_s^{min}), current (T_s) and maximum (T_s^{max}) step time,
- message composing time (T_c),
- encryption time (T_e),
- decryption time (T_d),
- time of generating confidential information (T_g),
- minimum (T_{ses}^{min}), current (T_{ses}) and maximum (T_{ses}^{max}) session time.

The distinction between minimum, current and maximum values is significant due to the specificity of the research. Thanks to this, we can determine the range of the tested network delay values and then indicate the specificity of the Intruder's behaviour in various time aspects. Also, we took into account the difference between symmetric and asymmetric encryption and decryption. Symmetric algorithms are faster than asymmetric, so we used different time value for both operations.

For the purposes of time analysis, we made the following assumptions:

- D_{min} = 1 [tu],
- D_{max} = 4 [tu],
- (symmetric) T_e = 4 [tu],
- (symmetric) T_d = 4 [tu],
- (asymmetric) T_e = 6 [tu],
- (asymmetric) T_d = 6 [tu],
- T_g = 1 [tu],
- T_c = 1 [tu].

Table 1. Summary of operations during first part of our protocol.

Step	**G**	**C**	**E**	D	!E	Min	Max
1	+	+	A	+	A	15	18
2	+,+	+	A,S	+	S	19	21
3	-	+	A	+	A	14	17
4	-	+	A	+	A	14	17

Source: self-elaboration

Table 1 shows the summary of operations executed during the first part of our protocol. Column **Step** contains step numbers. Columns **G**, **C** and **D** contain information about the occurrence of operations in the current step: message composing, generating confidential information and delay in the network respectively. Designation **+** means that such operation occurs in the current step, and designation **–** means that such operation does not occur in the current step. In this step, there DC generates confidential information two times, so we included T_g two times in our calculations. Columns **E** and **!E** contains information of algorithms used for encryption or decryption. Please note that in the second step of the first protocol part there DC encrypts two times. It uses both, asymmetric and symmetric encryption. In this same step, the receiver decrypts only symmetric. Columns **Min** and **Max** contain information of minimal and maximal step time.

Next, we calculated lifetimes for steps of our protocol first part:

- L_1= 73 [tu],
- L_2= 55 [tu],
- L_3= 34 [tu],
- L_4= 17 [tu].

Also, we calculated minimal and maximal session time:

- T_s^{min}= 73 [tu],
- T_s^{max}= 61 [tu].

Table 2. Summary of operations during second part of our protocol.

Step	**G**	**C**	**E**	D	!E	Min	Max
1	-	+	A	+	A	15	18
2	-	+	A,A	+	A	20	23
3	-	-	-	+	A	7	10
4	+	+	A	+	A	15	18

Source: self-elaboration

Table 2 shows the summary of operations executed the during second part of our protocol. Please note that there are two asymmetric encryption and only one asymmetric decryption. In the third step, user A sends a message which he received in the earlier step. So he does not generates confidential information and composes and encrypts the message in this step.

Next, we calculated lifetimes for steps of our protocol second part:

- L_1= 69 [tu],
- L_2= 51 [tu],
- L_3= 28 [tu],
- L_4= 18 [tu].

Also, we calculated minimal and maximal session time:

- T_s^{min} = 57 [tu],
- T_s^{max}= 69 [tu].

These values were necessary to enable and set time conditions.

For our protocol, the tool described in (Szymoniak, 2021) and (Szymoniak et al., 2021) generated seventy-two different executions. These executions were generated combinatorically. The difference between particular executions is the order users appear and the objects used by the Intruder. The generated executions can be divided into three types. The first is fair execution, in which only honest users are present. The second type of executions are executions with an Intruder who does not impersonate any user. The third type is executions, in which the intruder impersonates one of the honest users.

Next, we performed simulations of our protocol executions. For this part of the research, we used the randomly generated current delay in the network values. These values were randomly generated according to normal, uniform, Cauchy's and exponential probability distributions. We choose these probability distributions to model the real work of a computer network. Also, mentioned tool allows the generation of values out the accepted range $<D_{min}, D_{max}>$.

We will present the simulations of our protocol on normal and uniform probability distributions example. All executions were tested in thousand series. The simulations assume that execution can end with one of four statuses. The first status is *correct*. Here execution ends between T_s^{min} and T_s^{max}. The second status is !*min*. Here execution ends below T_s^{min} while the time conditions are met. The third status is !*max*. Here execution ends above T_s^{max} while the time conditions are met. The last status is *error*. Here execution ends with the failure to meet the time conditions.

For research with the first part of our protocol and normal probability distribution, executions ended with *correct*, *!max* end *error* statuses. We observed that only twenty executions were possible. This set consisted of six executions of the first type (numbered 1-6), fourteen executions of the second type (numbered 7-20). There were no executions of the third type, so the tool did not find an attack on our protocol.

Table 3 presents simulations results for these executions ended in the correct session time. We used the current delay in the network value generated according to a normal probability distribution. We present the average session time and the average delay in the network value. Please note that the session time depends on protocol execution. If Intruder does not need to get knowledge from an additional step, the session time is lower, and the execution can end with the correct session time.

Table 3. Executions ended in the correct session time.

No.	Average session time [tu]	**Average delay in the network [tu]**
1	61.61	2.90
2	61.94	5.64
3	61.96	4.06
4	62.43	4.05
5	62.52	4.27
6	62.91	4.87
7	63.04	3.55
8	63.51	4.35
9	64.22	3.17
10	64.42	3.51
11	64.52	4.96
12	64.72	4.34
13	65.79	4.25
14	65.82	5.63
15	65.97	5.22
16	66.62	5.88
17	66.72	4.61
18	66.77	5.43
19	67.01	5.61
20	67.05	15.61

Source: self-elaboration

Table 4 presents simulations results for possible executions ended above the correct session time. Also, here we checked the first part of our protocol with the current delay in the network value generated according to a normal probability distribution. Obtained results showed that additional steps executed by Intruder can increase the session time while the time conditions are met.

Table 4. Executions ended above correct session time.

No.	Average session time [tu]	**Average delay in the network [tu]**
1	82.03	10.20
2	82.04	10.20
3	82.05	10.21
4	82.05	10.01
5	82.06	10.61
6	82.06	10.21
7	82.07	10.01
8	82.08	10.61
9	82.11	10.02
10	82.13	10.22
11	82.14	17.47
12	82.14	10.02
13	82.15	10.62
14	82.17	10.03
15	82.22	10.24
16	82.23	10.04
17	82.23	10.04
18	82.25	10.05
19	82.28	14.35
20	82.30	10.45

Source: self-elaboration

Next, we perform research with a uniform probability distribution for the second part of our protocol. Similarly, for this stage of research, executions ended with *correct*, *!max* end *error* statuses. Also, we observed that only twenty executions were possible. This set consisted of six executions of the first type (numbered 1-6), fourteen executions of the second type (numbered 7-20).

Table 5 presents simulations results for these executions that ended in the correct session time. We used the current delay in the network value generated according to a uniform probability distribution. We observed a lower current delay in the network values than in the case of a normal probability distribution. These values affect session times and involve the Intruder to execute additional steps.

Table 5. Executions ended in the correct session time.

No.	Average session time [tu]	**Average delay in the network [tu]**
1	61.8	1.76
2	63.8	2.16
3	65.4	2.48
4	65.6	2.52
5	66.1	2.62
6	66.5	2.7
7	66.6	2.72
8	66.9	2.78
9	67.1	2.82
10	67.3	2.86
11	67.3	2.86
12	67.6	2.92
13	67.7	2.94

14	67.7	2.94
15	67.7	2.94
16	67.8	2.96
17	68	3
18	68	3
19	68.1	3.02
20	68.2	3.04

Source: self-elaboration

Table 6 presents simulations results for possible executions ended above the correct session time. We used the current delay in the network value generated according to a uniform probability distribution. We observed that session times were greater than maximal session time less than 1 [tu].

Table 6. Executions ended above correct session time.

No.	Average session time [tu]	**Average delay in the network [tu]**
1	69.1	5.94
2	69.1	7.2
3	69.2	8.76
4	69.2	6.74
5	69.2	5.5
6	69.3	6.22
7	69.3	6.74
8	69.4	6.56
9	69.4	6.42
10	69.4	6.2
11	69.5	7.04
12	69.5	6.38
13	69.5	7.3
14	69.5	7.22
15	69.6	7
16	69.6	7.66
17	69.6	5.74
18	69.6	6.94
19	69.6	7.68
20	69.7	7.1

Source: self-elaboration

CONCLUSION

This paper discussed the new security protocol for protection against false links. Security protocols are widely used for users’ protection in the virtual world. Their use gained importance during the COVID-19 pandemic, where many aspects of our lives had to move to the Internet. In the era of the development of cyber attacks, there was also a need to strengthen the security of user communication. This is where security protocols come in to help.

We presented the new security protocol which aims to users authentication and protection against false links. The proposed protocol consists of two parts: preparatory and verification. In the first part, the user must agree on his unique identifier with the trusted Distribution Center. Also, the event organizer must perform a similar action. In the second part, the user (who wants to participate in any event) will verify the correctness of the URL used to make the connection.

We used the tool described in (Szymoniak, 2021) and (Szymoniak et al., 2021) to check our protocol. We made timed simulations using the current delay in the network value generated according to selected probability distributions. We observed time influence on our protocol. Badly selected time dependencies could be met even the current session time is greater than maximal session time. Also, we observed that any execution with the Intruder who impersonates an honest user was not possible. This suggests that our protocol is secure.

In further work, we will focus on developing a URL verification method. As part of this method, we plan to build a method, which will indicate whether a given URL is false or good. Distribution Center will use this method. Ultimately, we plan to prepare a tool that will use the protocol and related methods to protect users from making a false link.

ACKNOWLEDGEMENTS

The project financed under the program of the Polish Minister of Science and Higher Education under the name "Regional Initiative of Excellence" in the years 2019 - 2022 project number 020/RID/2018/19, the amount of financing 12,000,000.00 PLN.

KEYWORDS: Security protocols, false URL detection, cybersecurity, neural network.

REFERENCES

Armando, A. & et. al. (2005). The AVISPA tool for the automated validation of internet security protocols and applications, In: Proc. of 17th Int. Conf. on Computer Aided Verification (CAV'05), vol. 3576 of LNCS, pp. 281-285, Springer.

Baccouche, A., Ahmed, S., Sierra-Sosa, D. & Elmaghraby, A. (2020). Malicious Text Identification: Deep Learning from Public Comments and Emails. Information, 11, 312.

Basin, D., Cremers, C. & Meadows, C. (2018). Model Checking Security Protocols, in Handbook of Model Checking, Springer International Publishing.

Blanchet, B. (2016). Modeling and Verifying Security Protocols with the Applied Pi Calculus and ProVerif, Foundations and Trends in Privacy and Security, vol. 1(1-2) pp.1–135.

Burrows, M., Abadi, M., & Needham, R. (1989). A Logic of Authentication, In: Proceedings of the Royal Society of London A, vol. 426.

Chadha, R., Sistla, P. & Viswanathan, M. (2017). Verification of randomized security protocols, Logic in Computer Science.

Čibej, U., Fürst, L. & Mihelič, J. (2019). A Symmetry-Breaking Node Equivalence for Pruning the Search Space in Backtracking Algorithms. Symmetry, 11, 1300.

Dolev, D. & Yao, A. (1983). On the security of public key protocols. In: IEEE Transactions on Information Theory, 29(2).

Galinec, D., Steingartner, W. & Zebić, V. (2019). Cyber Rapid Response Team: An Option within Hybrid Threats," 2019 IEEE 15th International Scientific Conference on Informatics, Poprad, Slovakia.

Ispahany, J., & Islam, R. (2020). Detecting Malicious URLs of COVID-19 Pandemic using ML technologies.

Ko, Y., Kim, J., Duguma, D.G., Astillo, P.V., You, I. & Pau, G. (2021). Drone Secure Communication Protocol for Future Sensitive Applications in Military Zone. *Sensors***.**

Kwon, D.K., Yu, S.J., Lee, J.Y., Son, S.H. & Park, Y.H. (2021). WSN-SLAP: Secure and Lightweight Mutual Authentication Protocol for Wireless Sensor Networks. *Sensors*.

Lai, C.-M., Shiu, H.-J. & Chapman, J. (2020) Quantifiable Interactivity of Malicious URLs and the Social Media Ecosystem. Electronics, 9.

Lowe G. (1996). Breaking and fixing the needham-schroeder public-key protocol using fdr. In Proceedings of the Second International Workshop on Tools and Algorithms for Construction and Analysis of Systems, TACAS ’96, pages 147–166, London, UK, 1996. Springer-Verlag.

Needham, R. M. & Schroeder, M. D. (1978). Using encryption for authentication in large networks of computers. Commun. ACM, 21(12).

Nigam, V. & et. al (2016). Towards the Automated Verification of Cyber-Physical Security Protocols: Bounding the Number of Timed Intruders, Computer Security – ESORICS 2016”, Springer International Publishing.

Maheswari, M., & Karthika, R.A. (2021). A Novel QoS Based Secure Unequal Clustering Protocol with Intrusion Detection System in Wireless Sensor Networks. *Wireless Pers Commun*.

Moreno-Cruz, F., Toral-López, V., Escobar-Molero, A., Ruíz, V.U., Rivadeneyra, A. & Morales, D.P. (2020). *treNch*: Ultra-Low Power Wireless Communication Protocol for IoT and Energy Harvesting. *Sensors*, *20*, 6156.

Patel, A. & Tailor, J. (2020). A malicious activity monitoring mechanism to detect and prevent ransomware, Computer Fraud & Security, Volume 2020, Issue 1, 2020, Pages 14-19.

Paulson, L. (1999). Inductive Analysis of the Internet Protocol TLS, ACM Transactions on Information and System Security (TISSEC), vol 2 (3).

Piątkowski, J. (2020). The Conditional Multiway Mapped Tree: Modeling and Analysis of Hierarchical Data Dependencies. IEEE Access 8: 74083-74092.

Radaković, D. & Herceg, D. (2018) Towards a completely extensible dynamic geometry software with metadata, Computer Languages, Systems & Structures, Volume 52, pp 1-20.

Siedlecka-Lamch, O. (2020). Probabilistic and timed analysis of security protocols, In proceeding of the 13th International Conference on Computational Intelligence in Security for Information Systems CISIS 2020, 16-18 September 2020, Burgos, Spain; paper 24.

Siedlecka-Lamch, O., Szymoniak, S. & Kurkowski, M. (2019) A fast method for security protocols verification, Computer Information Systems and Industrial Management, Springer.

Song, X., Chen, C., Cui, B. & Fu, J. (2020). Malicious JavaScript Detection Based on Bidirectional LSTM Model. Appl. Sci., 10, 3440.

Steingartner, W., Novitzka, V. & Schreiner, W. (2019). Coalgebraic operational semantics for an imperative language, Computing and Informatics, 38 (5), pp. 1181-1209.

Steingartner, W. & Galinec, D. & Kozina, A. (2021). Threat Defense: Cyber Deception Approach and Education for Resilience in Hybrid Threats Model, *Symmetry* 13, no. 4: 597.

Szymoniak, S. (2021). Security protocols analysis including various time parameters, Mathematical Biosciences and Engineering, 18(2): 1136-1153.

Szymoniak, S. (2020). How to be on time with security protocol?, Mario Arias Oliva, Jorge Pelegrín Borondo, Kiyoshi Murata, Ana María Lara Palma, Societal Challenges in the Smart Society ETHICOMP Book Series, Universidad de La Rioja, p. 225-237.

Szymoniak, S., Siedlecka-Lamch, O., Zbrzezny, A.M., Zbrzezny, A. & Kurkowski, M. (2021). SAT and SMT-Based Verification of Security Protocols Including Time Aspects. *Sensors*, *21*, 3055.

Xiao, D. & Jiang M. (2020). Malicious Mail Filtering and Tracing System Based on KNN and Improved LSTM Algorithm," 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Calgary, AB, Canada, 2020, pp. 222-229.

Yang, Z., Sun, Q., Zhang, Y. & Wang, W. (2020). Identification of Malicious Injection Attacks in Dense Rating and Co-Visitation Behaviors, in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 537-552, 2021, doi: 10.1109/TIFS.2020.3016827.

BETWEEN SCYLLA AND CHARYBDIS: THE "NEW NORMAL" CYBER RESILIENCE POSTURE OF CIVIL SOCIETY ORGANISATIONS

Mamello Thinyane, Christy Un, Debora Christine

United Nations University institute in Macau (Macau)

mamello@unu.edu; unchristy@unu.edu; debora@unu.edu

ABSTRACT

The COVID-19 pandemic has disrupted the normal trajectory of societal functioning across different sectors, including the third sector, and has forced the world to contend with the cascading impacts of what started as a localised disease outbreak now turned multisectoral global crisis. The role of digital technologies to support continuity and restoration of functioning during the pandemic has been critical, particularly for the civil society organisations which have taken up the growing burden to provide essential services to citizens – especially to vulnerable and marginalised populations. However, for most entities, including civil society organisations, digital technology has been a double-edged sword in that while it has supported resilience during the pandemic, it has also been a source of cyber risks, including cyberattacks and socio-technical threats, which have exacerbated the negative impacts of the crisis. This research investigates the cyber resilience posture of civil society organisations in the "new normal" and employs a resource dependency framework to analyse the effects of external environmental factors, that influence the organisational behaviour and structure, on the cyber resilience of the organisations. The research finds that the "new normal" cyber resilience posture is that civil society organisations remain in a precarious cybersecurity situation, where despite their services being more critical, their increased reliance on digital technologies, and their increased cyber threat exposure, they are still characterised by lack of financial resources - which is associated with lack of prioritisation of cybersecurity in funding instruments, skilled support, technical capacity, awareness of compliance risks, and ability to engage in long-term strategic and contingency cyber resilience planning. The paper advances the argument for prioritising and supporting civil society stakeholders as critical active agents towards the co-production of cyber resilience as a public good.

INTRODUCTION

The COVID-19 pandemic has disrupted the normal trajectory of societal functioning across different sectors, including the third sector, and has forced the world to contend with the cascading impacts of what started as a localised disease outbreak now turned multisectoral global crisis. The pandemic has highlighted the complex dependencies between different sectors and levels in society and foregrounded the value of societal resilience – the whole-of-society capability for positive adaptation during significant adverse incidents.

In response to the rise in demand for civic activism and action at the community level, civil society organisations (CSOs) have borne the brunt of the pandemic and assumed the growing burden to provide critical services to citizens, especially vulnerable and marginalised populations. CSOs worldwide have either filled up the vacuum left behind by government failure, partnered with government interventions, monitored government responses as state watchdogs, or advocated for structural reforms to the political, social, and economic models of the society (Youngs, 2020). The pandemic has shown civil society actors contributing to societal resilience through service provision,

social donations, information dissemination, and advocacy for marginalised populations (Cai, Okada, Jeong, & Kim, 2021).

The role of digital technologies to support continuity and restoration of functioning during the pandemic has been critical: schools have provided lessons online, businesses and organisations have shifted to virtual operations, and governments have digitised their services (Taddeo, 2020). For CSOs, the pandemic has catalysed new forms of civic mobilisation, which has seen organisations shifting to digital organising and increasing their collaboration with various stakeholders in emergency relief and informal activism (Brechenmacher, Carothers, & Youngs, 2020). Increasingly, these organisations also find the use of digital tools as a key mechanism to adapt to the "new normal" and for resource mobilisation to support their mission and goals (Miao, Schwarz, & Schwarz, 2021; Nemțeanu & Dabija, 2020; Youngs, 2020). A study by the European Economic and Social Committee found that 74% of CSOs reported reliance on digital technologies to implement new services and initiatives during the pandemic (European Economic and Social Committee, 2020).

However, while digital technologies have supported societal and CSOs' resilience during the pandemic, they have also increased exposure to adverse cyber incidents and risks. Congruent with the findings from the Global Risks Report 2021, the global digital transformation across sectors is associated with the increased global vulnerability and exposure to technological risks (The Global Risks Report 2021, 2021). Globally, there has been a surge in the number and range of COVID-19 related cyber-attacks because the pandemic has afforded threat actors new attack vectors, increased attack surface, and opportunities for social engineering attacks that capitalise on individuals' heightened psychological distress as well as on media and governmental messaging around the pandemic (Lallie et al., 2021). According to the latest research from Ecclesiastical Insurance, a third of the 250 charities surveyed have fallen victim to a cyber-attack during the pandemic. However, just over half (52%) of them had subsequently put in place a cybersecurity plan, and less than a quarter (23%) had enhanced investment in cybersecurity. This alludes to the challenges that CSOs face with regard to improving their cyber resilience (Ecclesiastical, 2020).

This paper centres on CSOs and highlights the challenges, contentions, and dilemmas surrounding their practice during the COVID-19 pandemic. First, the contentions inherent in the very notion of resilience in general and cyber resilience specifically are explicated. These contentions are then explored and elaborated through a case study of local CSOs. The paper posits and illustrates how the cyber resilience posture of CSOs has become more precarious and vulnerable during and, likely, post the COVID-19 pandemic - the "new normal". The following section presents the Resource Dependency Theory (RDT) framework used in this research to analyse the cyber resilience situation of CSOs. This is followed by the application of this framework for analysis of a case study of 35 CSOs to explore the environmental factors and critical resources affecting the cyber resilience of CSOs, as well as the role of organisations' management to manage the resource dependency, uncertainties, and constraints towards enhanced organisations cyber resilience. The paper concludes with a discussion of the key contentions that places CSOs in a position of continued cybersecurity precarity and vulnerability in the "new normal".

RESOURCE DEPENDENCY FRAMEWORK

Several approaches and theoretical frameworks can be employed to explore the cyber resilience situation of civil society organisations. From a technical cybersecurity perspective, the investigation can be framed around the maturity of the cybersecurity controls that the organisations are implementing across the various organisational domains, such as access control, asset management, training and capacity-building, and risk management and planning (Benz & Chatterjee, 2020; Carías, Arrizabalaga, Labaka, & Hernantes, 2020; Ross, Pillitteri, Graubart, Bodeau, & McQuaid, 2019). This

approach would be focused on cybersecurity as the locus of the analysis instead of the organisation, its behaviours, and structures, and how those contribute to the overall organisational cyber resilience. Further, this analysis would only reveal the performance and maturity of the organisation towards cyber resilience without exploring the complexities and contentions inherent in the cyber resilience situation of organisations. These complexities and contentions are associated with the environmental context in which the organisations operate, and they are manifested in the diverse decision-making and strategising that organisations' management undertakes towards cyber resilience.

This research focuses on the organisations as the locus of analysis and employs RDT, which frames organisational behaviour as dependent on the organisation's use of resources that ultimately emanate from its environment (Pfeffer & Salancik, 1978). RDT, as an open-systems approach to organisational analysis, recognises the influence of the external environment on the organisations' management (Seo, 2011). The theory links the behaviour, structure, and success of organisations to the management's decision-making abilities, which are influenced by the internal and external agents that have access, control, and leverage on critical resources (Nienhüser, 2008). The premise of RDT is that "the key to organisational survival is the ability to acquire and maintain resources", specifically critical resources (Pfeffer & Salancik, 1978, p. 2).

The notion of resource criticality in RDT is linked to the ability of organisations to continue functioning in the absence of a specific resource. Therefore, the more an organisation is dependent on specific resources for core functioning, the more the criticality of that resource (Nienhüser, 2008). For example, manufacturing organisations cannot function without the critical raw materials that go into the final products; service organisations likewise depend on critical input resources that contribute to the service delivery. Within RDT, resources are broadly defined to include "various assets, capabilities, organisational processes, information, and knowledge that contribute to improved organisational efficiency and effectiveness" (Seo, 2011, p. 3). RDT provides a formulation of the following basic elements (Nienhüser, 2008):

- The environment as a source of uncertainty and constraint – this gives recognition to the influence of the external environment in shaping both the access to critical resources and the decision-making and problem-solving ability of the organisational actors. Within the RDT, the environment is given primacy as the source of uncertainty for organisations (i.e., in contexts of limited critical resources) and as a basis for power (i.e., in contexts where organisations can concentrate critical resources and limit external dependencies). Invariably the environment also constrains the performance and activities of the organisations.
- The distribution of power within organisations and outside organisations – this builds on the hypothesis that whoever has control of critical resources has power over the actors who need the resources. This is true both in terms of the relationship with external stakeholders and the relationship between internal actors within an organisation. The role of management in brokering and facilitating this distribution of power, and the corollary, the uncertainty of the organisation, is neither wholly dictated by nor independent of the external environment. Management has several alternate actions around which they can frame their response and engagement with external stakeholders: through compliance (to the demands of the external actors), avoiding influence from the environment, avoiding dependence (by creating alternative resources), and by managing the social control by reducing the power and dominance of controlling actors. As far as the internal distribution of power within organisations is concerned, the RDT recognises that actants (i.e., organisational units, departments) that can cope with and are responsible for solving the organisation's critical resources end up acquiring more power within organisations. The theory also suggests that

internal units generally tend to operate and act towards consolidating their power within the organisation. This dynamic is observed not only in the decision-making and the organisational structures but also in the executive succession processes.

- The RDT recognises that decisions and actions within an organisation have a feedback effect that influences the subsequent use of resources within the organisation and the balance of inter and intra-organisational power.

The RDT is a sufficient theory for explaining the behaviour, structure, stability, and change of organisations (Nienhüser, 2008). It shares similarities with several other organisational theories, such as the Resource-Based View (RBV) theory. Both theories recognise the importance of (critical) resources towards the performance and success of organisations. However, the RDT emphasises the environment and external and internal power differentials that emanate from the critical resource scarcity. On the other hand, RBV focuses more on the internal organisational resources which contribute to organisational behaviour and success.

While the relevance of RDT is immediately apparent for private sector organisations, whose behaviour and activities are motivated by economic efficiency and profit-making goals, there has been a need to interrogate the relevance and sufficiency of the theory for explaining the behaviour of not-for-profit organisations, which are primarily motivated by missional goals. Verbruggen et al. (2009) studied how RDT provides a relevant framework for investigating not-for-profit organisations' compliance with financial reporting standards, which is associated with the demand for the organisations to be financially accountable and transparent (Verbruggen, Christiaens, & Milis, 2009). Seo (2011) showed the relevance of the RDT for explaining the success of not-for-profit organisations. The research specifically investigated how the Resource Dependence Patterns (RDPs) of resource dependency, resource diversity, resource uncertainty, resource abundance, and resource competitiveness affected the behaviour, performance, and survival of not-for-profit organisations. Further, concerning the not-for-profit sector, RDT has been considered with respect to how it has transformed the sector to be more commercialised and to adopt private sector techniques, including competitive consolidation of resources, which may not necessarily be conducive with the values and traditional *modus operandi* of not-for-profit organisations, and which in the long run can lead to the reduced quality of services (Curley, Levine Daniel, Walk, & Harrison, 2021; Entwistle & Martin, 2005; Randle, Leisch, & Dolnicar, 2013).

RDT has been criticised for overly relying on the formulation of politics (i.e., the power that emanates from the control of critical resources) for framing organisational behaviour and activities. Other perspectives, including economic and instrumental efficiency approaches, which also provide an adequate explanation of organisational behaviour, have generally been neglected within the RDT.

Despite these limitations and critiques, the RDT is adopted in this paper as a theoretical framework for unpacking the resilience, specifically cyber resilience, of not-for-profit and civil society organisations. Resilience, which has its roots in the Latin word "resilire", means "to jump back, recoil". Therefore, implicit in the notion of resilience is the capability to bounce back and restore normal functioning after disturbances – hence the definition of resilience as the probability of persistence (Holling, 1973). Resilience has been formulated from different domains, including psychology, engineering, ecological studies, and disaster studies (Bourbeau, 2013). Socio-ecological resilience extends the framing of resilience to emphasise the achievement of an improved position after the disturbance, thereby introducing the notion of "bouncing forward" (Chandler, 2019). For organisations, resilience is traditionally formulated from the engineering and organisational psychology perspectives, but more

recently conceptualised as complex system survival towards the goal of maintaining critical system functions and processes, as opposed to returning to equilibrium or achieving system efficiency (Mamouni Limnios, Mazzarol, Ghadouani, & Schilizzi, 2014). This perspective is congruent with RDT as far as the critical system functions are dependent on critical resources.

Across different domains, resilience calls for diverse stakeholders' participation – due to its systemic nature and for engagement with key tensions and contentions associated with resilience practice. For example, it calls for engagement with the politics of delineating system boundaries, such as who is included and excluded; the complexity of formulating persistence and adaptation criteria; the value-laden acknowledgement of the non-normativity of resilience as well as the emancipatory catastrophism effects of crises (Beck, 2015; Mamouni Limnios et al., 2014).

As far as organisations are concerned, cyber resilience is enacted and achieved through the employment of various critical cybersecurity-related resources that organisations have access to; this is explained by and linked to the notion that resource dependency strategies influence organisational resilience (Roundy & Bayer, 2019). Further, while the cybersecurity risk exposure of organisations can be associated with internal risks, most of the cybersecurity risks that organisations deal with emanate from the external environment. The RDT allows for a systematic exploration and articulation of these dynamics that affect the overall cyber resilience of organisations.

RESEARCH DESIGN AND METHODOLOGY

The project is framed as a case study that employs convergent (i.e., concurrent) design mixed-method approach to explore the cyber resilience situation of civil society organisations and to cross-validate the findings (Fetters, Curry, & Creswell, 2013).

Thirty-five (35) CSOs in Macau SAR were engaged through an online survey and semi-structured interview instruments to investigate their cyber resilience posture along the following lines of inquiry: how digital resources have supported their resilience during the pandemic, their organisational cybersecurity situation and posture, and their cybersecurity threat exposure and experiences. Primarily a convenience sampling approach was used to recruit participants from the organisations with which that the researchers had a prior working relationship. A chain referral approach was also used to recruit further participants.

The analysis of the quantitative survey instrument data was undertaken through basic descriptive statistics identifying the situation of the CSOs at the aggregate level across the various lines of inquiry. Template analysis was used for the qualitative data, with the *a priori* codes and themes informed by the common cybersecurity domains and controls (i.e., the NIST CSF and the CIS domains).

ENVIRONMENTAL UNCERTAINTIES AND CONSTRAINTS

The investigation of the civil society organisations' cyber resilience, which is framed from the resource dependency perspective, first explores the environmental context of the civil society organisations. This unpacks the interactions between the organisations and their environment to highlight the uncertainties and constraints that emanate from the environment and that affect the cyber resilience of the CSOs. However, while the RDT primarily frames the dependency on external critical resources as a source of uncertainty, risk, and a loss of power, the potential of external resources to provide support and increase the resilience of organisations is recognised and explored in this analysis.

The general situation and environmental context of CSOs worldwide, which has been well-documented in the literature, is that they are generally under-resourced, operate in increasingly complex regulatory and compliance environments, tend to be marginalised in the cybersecurity domain and operate in an evolving cybersecurity risk environment (Brooks, 2020; Crete-Nishihata et al., 2014; Franz, Hayes, & Hannah, 2020; Jagalur et al., 2019).

The impact of the COVID19 pandemic

The single most significant phenomenon that has recently affected CSOs globally is the ongoing coronavirus pandemic, which has impacted the organisations directly and had cascading impacts that have indirectly affected CSOs. One of the direct impacts of the pandemic to the CSOs is that it has put the organisations, including their service users, at more significant health risks. This is because most CSOs have a socially-focused mission of providing services to specific civil society communities and stakeholders, which implies close contact and engagement with people and the increased potential for infections to propagate. Most (i.e., 79%) of the CSOs engaged in this research provide social welfare services and the most common (i.e., 30%) service clients across the organisations are the elderly.

The increased risk profile of the elderly means that most CSOs have also needed to alter their operations to cater to this increased risk level and respond to the mandated containment and isolation measures. In fact, in some countries, the highest cases of infections and fatalities have been clustered around social services centres catering to the elderly.

In recounting the impact that the pandemic has had on their operations and reflecting on how digital technology has enabled continued social interaction with service clients' families, despite the lockdown, one of the directors noted that:

> *"Because now we have the Covid-19, they were not allowed to enter inside. So, our social workers, they [sic] use these different methods that they may talk with video calls and talk to the family."*

This snippet sheds light on the impact of the pandemic at the grassroots and community level and highlights how the organisations have mitigated this challenge and put in place measures to address the impact of this adverse external phenomenon. The interview snippet above also sheds light on how the strategic mobilisation of digital resources and reengineering of processes and procedures enabled business continuity and resilience for the organisations. Organisations adopted digital technologies to facilitate continuity of their operations and facilitate interaction with their service users and families. For example, some organisations continued providing counselling services to students at schools using digital technologies. Further, to deal with social distancing measures and limit outside public activities, some organisations introduced computer-based interactive activities for their service users. However, while the use of digital technologies has been catalysed by and increased during the pandemic, in general, the CSOs are increasingly dependent on ICT for their daily operations at varying levels, with 11% and 60% indicating being "very reliant" and "reliant" respectively (Figure 1).

The second significant way that the organisations have been affected by the pandemic is through the secondary effects and cascading impacts from the environment. This is because CSOs engage with and depend on myriad other external stakeholders both as part of their existence and operations. The stakeholders that organisations engage with range from *ad hoc* service providers, such as repairs and maintenance service providers, to the core and critical "partners", such as funders and affiliate

organisations. From the resource dependency perspective, this dependence on external stakeholders – ultimately their resources, amplifies the power differential between the organisations and the external stakeholders. This dependency is also framed as the primary source of uncertainty for organisations.

Figure 1. Civil Society Organisations' reliance on ICT.

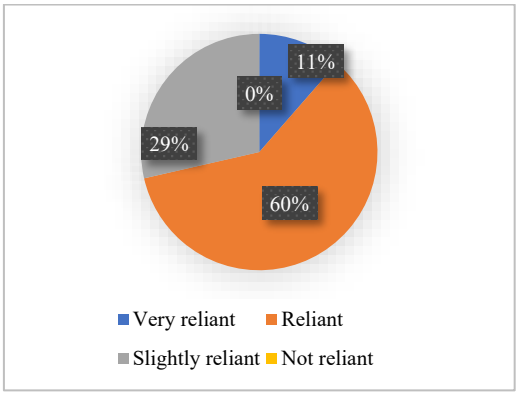
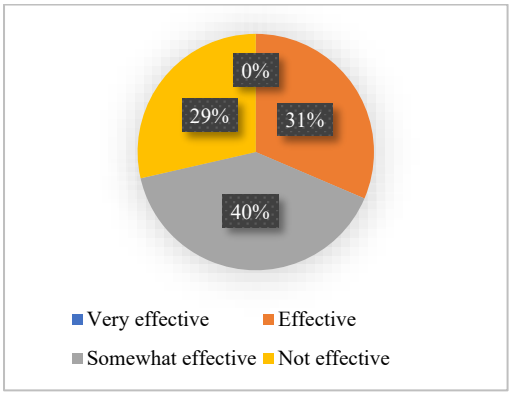


Figure 2. Effectiveness of CSOs cybersecurity policies.



In the case of the CSOs engaged in this project, one of the significant uncertainties has been manifested through funding and budget constraints. Most of the CSOs receive their funding from the social welfare bureau of the government, and in general, this is typically generous funding that enables them to effectively provide their services without financial limitations. However, because of the ongoing pandemic, the government has instituted measures to reduce expenditure across different sectors, which has affected the operations of the CSOs. The limited financial resources mean that CSOs have less money to spend on what is considered non-core and non-missional expenditures, such as cybersecurity-related expenditures.

The third indirect impact of the COVID19 pandemic on the organisations is associated with increased COVID19-related cybersecurity risks. Worldwide, the pandemic saw many organisations and businesses switching operations to the virtual mode, which led to an increased burden on digital infrastructure and, in some cases, to reduced quality of digital services. Further, the working from home modality also meant that the security cocoon that is provided within organisations networks was compromised, both by potentially unprotected personal devices connecting to the organisations' networks and by organisations' devices and data, being taken out of the protected organisations' perimeter. The pandemic created many opportunities for cybersecurity threat actors to exploit the new attack vectors and increased attack surface. Not surprisingly, with the onset of the pandemic, an upshoot in the number of COVID19 related social engineering attacks, including phishing, misinformation, and disinformation was observed worldwide. While these attacks were not exclusively targeted at CSOs, nor CSOs being the only sector experiencing these risks, they still represent one of the highly targeted sectors, making up 32% of victims of nation-state attacks observed between July 2019 and June 2020 (Microsoft, 2020).

Dependencies and interactions with external stakeholders

One of the key observations within the RDT regards the interaction of organisations with their external stakeholders, that "organisations will tend to be influenced by those who control the resources they

require" (Pfeffer & Salancik, 1978, p. 44). As far as organisational cyber resilience is concerned, several dependencies can be observed that affect the organisations' behaviour and operations and their cyber resilience posture.

The primary external dependency for CSOs worldwide is on funders who provide the critical financial resources for the organisations to exist and operate. As noted previously, this dependency is also one of the key sources of uncertainty and constraints for organisations in terms of the expenditure and reporting requirements associated with the funding.

Most of the organisations engaged in this research receive their funding from the government, which provides guidelines for the kinds of expenditure that organisations can budget for. The extent that the funding stakeholders recognise the importance of cybersecurity for CSOs, is the extent to which the organisations can prioritise cybersecurity investments in their budgets. Our research has observed that the current funding guidelines from the local government do not make explicit allowance for cybersecurity expenditure for the organisations. Notably, the Social Service Facilities' Regular Funding Budget Guidelines does not mention investment in cybersecurity or digital technology, except for the procurement and disposal of fixed property from public departments, private donors, and individual organisations. Furthermore, cybersecurity is not indicated in the section on organisational management mechanisms, which cover the management of organisations operations, human resources, finance, and reputation, except for a call for organisations to establish a guideline on the use and protection of sensitive data.

The second key dependency on external stakeholders is for information technology and cybersecurity resources and services. Due to limited internal resources, for example, that globally most CSOs spent approximately 5.7% of their budget on ICT or that, in general, they have a much lower ratio of technical staff than their private sector counterparts, CSOs rely on IT service providers for the procurement and maintenance of their IT resources (Brooks, 2018; Hulshof-Schmidt, 2017). They also need the services of computer emergency response teams (CERTs) who have the expertise to provide cybersecurity incident handling; this is despite most of the organisation indicating that they handle adverse cyber incidents internally (discussed in later sections).

Organisations also must deal with the dependencies associated with cybersecurity legislative compliance requirements, for example, with regards to personal data protection and disclosure of data breaches. This is because CSO's deal with personally identifiable and sensitive data of their service clients. Worldwide, CSO's are increasingly collecting social indicators data both in compliance with the requirements of the funders and to improve their services. At the same time, most jurisdictions around the world, including the local context, are putting in place personal data protection legislation that all data processors, including CSO's, must comply with.

ORGANISATIONAL MANAGEMENT OF CYBER RESILIENCE

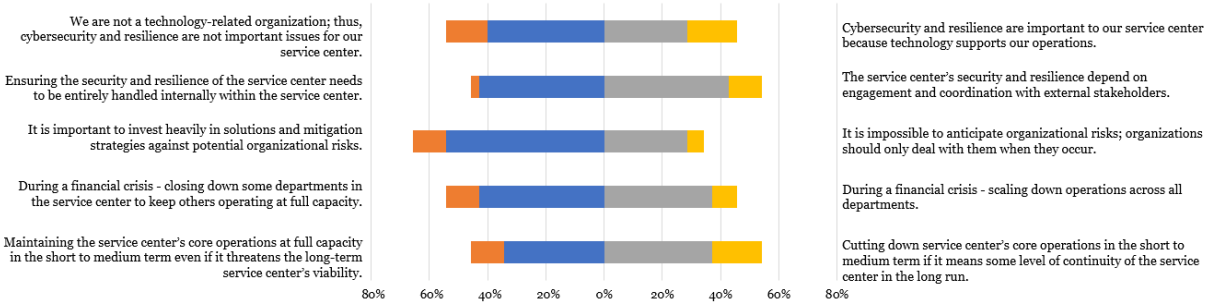
From the resource dependency perspective, the role of management towards organisational cyber resilience should be understood as "neither completely determined by the environment" nor comprising of the management's independent realisation of organisational goals and strategies (Nienhüser, 2008, p. 14). Management's key function is to manage and negotiate the organisations' interaction with the external environmental factors and their response to the demands of the external stakeholders, be it the funders, partners, or service users. For most organisations, service users are one of the key stakeholders that they identify as critical for their existence and continued operation. The organisations identify the ability to continue providing services to their clients as one of the critical

goals of resilience and, therefore, their persistence criteria, followed by the ability to minimise the impact of adverse incidents.

Beyond managing the interaction with external stakeholders, management's function is also to manage the internal organisational allocation of resources and the associated distribution of power between the different internal stakeholders, actors, and organisational units. Invariably, the stakeholders who handle the most critical operations within an organisation tend to gain more power and influence over other stakeholders. However, even with respect to the management of internal resources and processes towards cyber resilience goals, there remain several contentions that CSOs are confronted with regarding the outworking of these resilience goals. These contentions are primarily associated with the need for management to choose between divergent resource dependency strategies.

Illustratively, when presented with orthogonal organisational adaptation choices, the pathways towards the organisational resilience goal are varied between the different CSO managers and directors. For example, 51% of the CSOs' managers and directors would opt for maintaining operations at full capacity in the short-term (i.e., at the expense of long-term survivability) versus 48% that would opt to cut down operations to ensure long-term continuity; 54% of CSOs would opt to close some departments to keep others running at full capacity during a financial crisis versus 46% who would opt to scale down operations across all departments for the same crisis scenario (Figure 3).

Figure 3. Resource dependency strategies and choices towards organisational resilience.



The complex demands for the management of organisational cyber resilience mean that the skills, understanding, and perceptions of the organisations' management are critical. For the organisations that participated in this research, the manager and directors generally have limited cybersecurity management capacities regarding the risk management competencies, cybersecurity risks awareness, and the understanding of the local cybersecurity landscape. The limited ability of CSO's management to accurately perceive the risks and understand the critical cybersecurity resources and actors hampers their ability to reduce organisational uncertainty and enhance organisational cyber resilience.

DISCUSSION

Civil society organisations are finding themselves increasingly relying on digital technology, not only to deal with the COVID19 pandemic but also as part of the organisational evolution towards digitisation. However, for most CSOs, digital technology has been a double-edged sword in that while it has supported resilience during the pandemic, it has also been a source of cyber risks, including cyberattacks, socio-technical risks, and compliance risks, which have further exacerbated the negative

impacts of the crisis (Weil & Murugesan, 2020). The digital technologies, which in the case of CSOs are supplied and maintained by external IT service providers, increase the dependency of the CSOs on these external stakeholders with the corollary of also increasing the uncertainty and risks for the CSOs.

The ability of the organisations to reduce this uncertainty is primarily constrained by their limited access to funding resources and the demands of funding actors. Even for cases where the managers and directors of the organisations believe there is a need for cybersecurity investments, they are limited by the external funding guidelines and directives that do not prioritise non-core cybersecurity expenditure. One of the CSO's managers highlighted this constraint that they must deal with whenever they need to acquire or replace IT equipment, as follows:

> *'So, we have to apply to replace some hardware facilities, because we have been using the same system for a very long time, that is, the same hardware. Those [sic] hardware may have been products from eight years ago, or it was eight years ago when we bought it. Production is not an issue of how many years ago. So, you know that these technological matters vary on a monthly and yearly basis, right? We all know that our hardware equipment may have a huge gap with the new equipment nowadays. right? So, these are also the feedback we received from the IT companies. If we change that system, it should actually fasten [sic] some of our speed and be relatively stable.'*

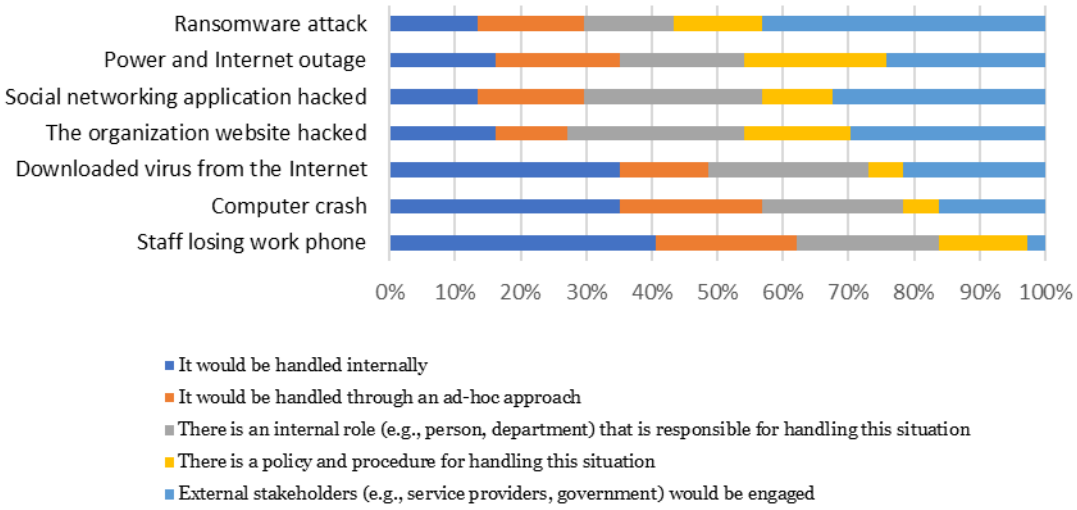
In the above snippet, despite the awareness of the risks of the old and obsolete equipment and the advantages of upgrading the equipment, the manager is constrained by the requirements of the external funder. Further, despite the management being aware of cyber resilience gaps within their organisations, they are limited from addressing those gaps due to limited resources and capabilities. For example, 29% and 40% of the managers noted that their cyber resilience policies were "not effective" and "somewhat effective", respectively (see Figure 2) and emphasised the need for further resourcing and capacity building for the managers as follows.

> *'I believe that in fact, the best way is to continue to provide training to us... Not only our center, but our entire association also needs some training to let more employees or service unit directors realise how important this matter is, and the profound impact it has.'*

Therefore, in general, CSOs remain financially constrained and under-resourced, and consequently in a weak cyber resilience position which subsequently increases their vulnerability and hampers their overall organisational resilience (Franz et al., 2020; Jagalur et al., 2019).

Given the importance of cyber resilience for CSOs' functioning, the global neglect of CSOs within the cybersecurity domain exacerbates their cybersecurity risk exposure and vulnerability. Research has found that threat intelligence reporting has led to an understatement of the impact of adverse cyber events on civil society, thus exposing CSOs to more vulnerabilities and threats in cyberspace (Maschmeyer, Deibert, & Lindsay, 2020). This neglect of CSOs, within the cybersecurity ecosystem further extends to the dearth of Computer Emergency Response Teams (CERTS), who provide incident handling support for CSOs. This situation leaves CSOs in a precarious and vulnerable position where they are forced to internally effect their cybersecurity solutions and handle cybersecurity incidents on an *ad-hoc* basis (Figure 4).

Figure 4. CSO's handling of cybersecurity incidents.



However, this similitude of internal capacity to handle adverse cyber incidents and the independence of the CSOs masks the sad and precarious reality that CSOs handle adverse cyber incidents internally because of limited resources to engage external service providers. The internal incident handling posture is not because of the availability of internal capacity because at most, only 11% of the CSOs in this research indicated having dedicated IT personnel within the organisation.

This situation highlights the need for several interventions towards enhancing the cyber resilience of CSOs: at the environmental level, the need for mainstreaming CSOs in national cybersecurity strategies, programs, and funding instruments; at the organisational management level, the need for capacity-building and resourcing to improve the decision-making, negotiation, and stakeholder-engagement capabilities of the management. Ultimately, there is a need to consider the complex dependencies and multistakeholder interactions that not only contribute to the cyber resilience of CSOs but to civil society and national cyber resilience as well.

CONCLUSION

The ongoing pandemic has impacted the normal functioning of society as we know it and ushered in the ensuing era of the "new normal", which is more reliant on digital technologies, exposed to more technological risks, and with complex interconnections and dependencies across stakeholders, levels, sectors, and domains. As far as the CSOs are concerned, the current "new normal" is that they remain in a precarious cybersecurity situation, where despite their services being more critical, their increased reliance on digital technologies, and their increased cyber threat exposure, they are still characterised by lack of financial resources which is associated with lack of prioritisation of cybersecurity in CSOs funding instruments, skilled support, technical capacity, awareness of compliance risks and ability to engage in long-term strategic and contingency planning (Brooks, 2020; Crete-Nishihata et al., 2014; Franz et al., 2020; Jagalur et al., 2019).

CSOs have to resort to employing frugal, creative, and collaborative solutions to their cyber resilience challenges (Huang & Pearlson, 2019). Notwithstanding the need for cyber resilience practices and cybersecurity solutions to be framed across the different layers of cyberspace, it is imperative to consider civil society not only as the most critical and vulnerable (i.e., linked to the notion human-

centric perspective to cybersecurity) dimension of the cyber ecosystem but also an active contributor to the co-production of cyber resilience (Gioe, Goodman, & Wanless, 2019).

ACKNOWLEDGEMENTS

This work is supported by the Science and Technology Development Fund of Macau (FDCT) under Grant No. 0016/2019/A.

KEYWORDS: Resilience, Cyber Resilience, Civil Society Organisations, Cybersecurity.

REFERENCES

Beck, U. (2015). Emancipatory catastrophism: what does it mean to climate change and risk society? *Current Sociology, 63*(1), 75–88.

Benz, M., & Chatterjee, D. (2020). Calculated risk? A cybersecurity evaluation tool for SMEs. *Business Horizons*, *63*(4), 531–540. https://doi.org/10.1016/j.bushor.2020.03.010

Bourbeau, P. (2013). Resiliencism: premises and promises in securitisation research. *Resilience*, *1*(1), 3–17. https://doi.org/10.1080/21693293.2013.765738

Brechenmacher, S., Carothers, T., & Youngs, R. (2020). *Civil Society and the Coronavirus : Dynamism Despite Disruption*. Retrieved from https://carnegieendowment.org/2020/04/21/civil-society-and-coronavirus-dynamism-despite-disruption-pub-81592

Brooks, S. (2018). *Defending Politically Vulnerable Organisations Online*. Retrieved from https://cltc.berkeley.edu/defendingpvos/

Brooks, S. (2020). *Digital Safety Technical Assistance at Scale*. Retrieved from https://cltc.berkeley.edu/security-at-scale/

Cai, Q., Okada, A., Jeong, B. G., & Kim, S. J. (2021). Civil society responses to the COVID-19 pandemic: A comparative study of China, Japan, and South Korea. *China Review*, *21*(1), 107–137.

Carías, J. F., Arrizabalaga, S., Labaka, L., & Hernantes, J. (2020). Cyber resilience progression model. *Applied Sciences (Switzerland)*, *10*(21), 1–32. https://doi.org/10.3390/app10217393

Chandler, D. (2019). Resilience and the end(s) of the politics of adaptation. *Resilience*, *7*(3), 1–10. https://doi.org/10.1080/21693293.2019.1605660

Crete-Nishihata, M., Dalek, J., Deibert, R., Hardy, S., Kleemola, K., McKune, S., … Wiseman, G. (2014). *Communities @ Risk: Targeted Digital Threats Against Civil Society*.

Curley, C., Levine Daniel, J., Walk, M., & Harrison, N. (2021). Competition and Collaboration in the Nonprofit Sector: Identifying the Potential for Cognitive Dissonance. *Administration & Society*, 009539972110058. https://doi.org/10.1177/00953997211005834

Ecclesiastical. (2020, October 19). A third of charities have suffered a cyber-attack during the coronavirus pandemic. Retrieved May 10, 2021, from https://www.ecclesiastical.com/media-centre/third-of-charities-had-cyber-attack-during-pandemic/

Entwistle, T., & Martin, S. (2005). From competition to collaboration in public service delivery: A new agenda for research. *Public Administration*, *83*(1), 233–242. https://doi.org/10.1111/j.0033-3298.2005.00446.x

European Economic and Social Committee. (2020). *The response of civil society organisations to face the COVID-19 pandemic and the consequent restrictive measures adopted in Europe*. Retrieved from https://www.eesc.europa.eu/en/our-work/publications-other-work/publications/response-civil-society-organisations-face-covid-19-pandemic-and-consequent-restrictive-measures-adopted-europe-study

Fetters, M. D., Curry, L. A., & Creswell, J. W. (2013). Achieving integration in mixed methods designs - Principles and practices. *Health Services Research*, *48*(6 PART2), 2134–2156. https://doi.org/10.1111/1475-6773.12117

Franz, V., Hayes, B., & Hannah, L. (2020). *Civil Society Organisations and General Data Protection Regulation Compliance*.

Gioe, D. V., Goodman, M. S., & Wanless, A. (2019). Rebalancing cybersecurity imperatives: patching the social layer. *Journal of Cyber Policy*, *4*(1), 117–137. https://doi.org/10.1080/23738871.2019.1604780

Holling, C. S. (1973). Resilience and Stability of Ecological Systems. *Annual Review of Ecology and Systematics*, *4*(1), 1–23. https://doi.org/10.1146/annurev.es.04.110173.000245

Huang, K., & Pearlson, K. (2019). For What Technology Can't Fix: Building a Model of Organizational Cybersecurity Culture. *Proceedings of the 52nd Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2019.769

Hulshof-Schmidt, R. (2017). *The 10th Annual Nonprofit Technology Staffing and Investments Report*. Retrieved from https://www.nten.org/wp-content/uploads/2019/11/2017-Nonprofit-Technology-Staffing-and-Investments-Report_updated-2019.pdf

Jagalur, P. K., Levin, P. L., Brittain, K., Dubinsky, M., Landau-Jagalur, K., & Lathrop, C. (2019). Cybersecurity for civil society. *International Symposium on Technology and Society, Proceedings*, *2018-Novem*, 102–107. https://doi.org/10.1109/ISTAS.2018.8638270

Lallie, H. S., Shepherd, L. A., Nurse, J. R. C., Erola, A., Epiphaniou, G., Maple, C., & Bellekens, X. (2021). Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic. *Computers and Security*, *105*, 102248. https://doi.org/10.1016/j.cose.2021.102248

Mamouni Limnios, E. A., Mazzarol, T., Ghadouani, A., & Schilizzi, S. G. M. (2014). The resilience architecture framework: Four organisational archetypes. *European Management Journal*, *32*(1), 104–116. https://doi.org/10.1016/j.emj.2012.11.007

Maschmeyer, L., Deibert, R. J., & Lindsay, J. R. (2020). A tale of two cybers - how threat reporting by cybersecurity firms systematically underrepresents threats to civil society. *Journal of Information Technology & Politics*, 1–20. https://doi.org/10.1080/19331681.2020.1776658

Miao, Q., Schwarz, S., & Schwarz, G. (2021). Responding to COVID-19: Community volunteerism and co-production in China. *World Development*, *137*. https://doi.org/10.1016/j.worlddev.2020.105128

Microsoft. (2020). *Microsoft Digital Defense Report*. Retrieved from https://download.microsoft.com/download/f/8/1/f816b8b6-bee3-41e5-b6cc-e925a5688f61/Microsoft_Digital_Defense_Report_2020_September.pdf

Nemțeanu, M.-S., & Dabija, D. (2020). Best Practices of nongovernmental organisations in combatting Covid-19. In R. Pamfilie, V. Dinu, L. Tăchiciu, D. Pleșea, & C. Vasiliu (Eds.), *New Trends in Sustainable Business and Consumption* (pp. 626–633).

Nienhüser, W. (2008). Resource dependence theory: How well does it explain behaviour of organisations? In *Management Revue* (Vol. 19). Retrieved from Rainer Hampp Verlag website: http://hdl.handle.net/10419/78991

Pfeffer, J., & Salancik, G. (1978). *The External Control of Organisations: A Resource Dependence Perspective*. Harpercollins College Div.

Randle, M., Leisch, F., & Dolnicar, S. (2013). Competition or collaboration? The effect of non-profit brand image on volunteer recruitment strategy. *Journal of Brand Management*, *20*(8), 689–704. https://doi.org/10.1057/bm.2013.9

Ross, R., Pillitteri, V., Graubart, R., Bodeau, D., & McQuaid, R. (2019). NIST Special Publication 800-160 Volume 2 - Developing cyber resilient systems: *NIST Special Publication*, *2*. https://doi.org/10.6028/NIST.SP.800-160v2

Roundy, P. T., & Bayer, M. A. (2019). To bridge or buffer? A resource dependence theory of nascent entrepreneurial ecosystems. *Journal of Entrepreneurship in Emerging Economies*, *11*(4), 550–575. https://doi.org/10.1108/JEEE-06-2018-0064

Seo, J. (2011). *Resource Dependence Patterns and Organisational Performance in Nonprofit Organisations*. Arizona Sate University.

Taddeo, M. (2020, June 1). The Ethical Governance of the Digital During and After the COVID-19 Pandemic. *Minds and Machines*, Vol. 30, pp. 171–176. https://doi.org/10.1007/s11023-020-09528-5

The Global Risks Report 2021 (16th ed.). (2021). Retrieved from http://wef.ch/risks2021

Verbruggen, S., Christiaens, J., & Milis, K. (2009). *Can resource dependence explain not-for-profit organisations' compliance with reporting standards?* (No. 2009/24).

Weil, T., & Murugesan, S. (2020). IT Risk and Resilience-Cybersecurity Response to COVID-19. *IT Professional*, *22*(3), 4–10. https://doi.org/10.1109/MITP.2020.2988330

Youngs, R. (2020). *Global Civil Society in the Shadow of Coronavirus*. Retrieved from https://carnegieendowment.org/files/Youngs-Coronavirus_Civil_Society_final.pdf

TECHNOLOGY AND GEOECONOMICS: EMERGING CONFLICTS IN THE DIGITAL WORLD

Nehme Khawly, Mario Arias-Oliva, Jorge De Andres

Universitat Rovira i Virgili (Spain), Complutense University of Madrid (Spain), Universitat Rovira i Virgili (Spain)

nehme.alkhawly@estudiants.urv.cat; mario.arias@ucm.es; jorge.deandres@urv.cat

ABSTRACT

Environmental analysis is a key area of strategic Marketing. The new global environment is changing their nature during the last decades, emerging new concepts such as the digital revolution and the geoeconomics' challenges. The nature of international conflicts is changing, shifting from the diplomatic and military arena to the technological and economic one. Nowadays, telecommunications warfare is standing up strong as the protagonist of the current scenario. Superpowers are competing on who will be playing that role on the international scene. The fifth generation of telecommunications (5G) is pushing China and USA in a geoeconomics conflict. Both countries are fighting to set up their best strategies of propagating their geoeconomic influence and attaining their strategic goals, preserving their digital sovereignty from any kind of abuse and violation.

This paper focuses on the geoeconomic characteristics of the 5G warfare, sheds lights on its manifestations in the recent emerging events on the international scene, explains the way in which China and USA are dealing with it through displaying methods of cyber-diplomacy implemented by these powers to achieve their aims, and illustrates the repercussions of that techno-war at the politico-economic level, and most importantly at the level of individuals' health and their immune system.

INTRODUCTION

A non-armed war, is occurring nowadays all over the planet: which country will be the leader of the 5G innovation in the internet and telecommunication technology? A Napoleonic remark uttered more than 200 years ago stated: *"Let China Sleep, for when she wakes, she will shake up the world"*, and it seems that 5G has waken China up, and the world started to shake.

Indeed, the world is living now in a transitional phase, and preparing itself to move from the 4G to the 5G stage… And following is a brief of the past stages from 1G to 4G, which have paved the path to their successor: 5G.

The first generation of mobile telecommunications, 1G allowed us to make wireless phone calls; 2G was the tech that enabled texting options; 3G brought web browsing; and 4G made the video-streaming come true, and insured a constant connection with GPS satellites, which has permitted the rise of many companies like Uber (Bremmer, 2019).

The current phase we are living in, is the transitional one, for we are shifting from 4G to 5G, which will be totally different than the previous upgrades known at the telecommunications and technology levels. And here comes the chaos, USA is against the Chinese integration in their infrastructures and network through – the largest provider of telecom equipment in the world – Huawei's equipment and hardware. Main arguments are:

- For fear of China's spying the American companies' database because they will be streamed through their own hardware;
- Using Huawei equipment could leave the US susceptible to China's infrastructure attack, if they ever went into war (Bremmer, 2019).

Thus, the American administration spread political and economic tension between its European allies, to limit Huawei's participation in building 5G networks, to remove all its equipment from their markets, and to end up all the agreements done between their companies and mobile network providers from a side, and Huawei's company from the other.

Does this 5G constitute a real threat to the cyber sovereignty of the USA? And is it really capable of spying and executing cyber attacks against its database, and violates its security? It is true that as technology evolves and progresses, the information and evidence become more susceptible to abuse, but is that the main goal behind this emerging war between China and USA? Breaching each other's cyber sovereignty?

In the 2020 Annual Report to Congress *of the* U.S. – China ESRC, it was clearly stated, in the key findings, that China is consistently working on a new model and standards for the global order: "The Chinese Communist Party (CCP) seeks to revise the international order to be more amenable to its own interests and authoritarian governance system. It desires for other countries not only to acquiesce to its prerogatives but also to acknowledge what it perceives as China's rightful place at the top of a new hierarchical world order" (U.S. – China ESRC, 2020, p. 80).

Furthermore "The Chinese government views technical standards as a policy tool to advance its economic and geopolitical interests" (U.S. – China ESRC, 2020, p. 81).

After all, it seems that 5G is the new bait, to attract and catch the global market, and Huawei is the most competitor bait-maker. But is this lure going to catch only good preys of the market? What about dangerous ones? What if it fails to achieve its target? What if it backfired and gets opposite impacts to the ones China and Huawei initially intended?

Many questions and queries are treated in this study and are waiting to be explored.

METHODOLOGY

This paper addresses the 5G from two main perspectives: the politico-economic and the technological. In order to treat this subject and answer all its questions, we are going to analyze secondary data by adopting the narrative and analytical methodologies. Therefore, data is collected from books, academic journals, and newspapers, in addition to readings from global and valuable economic, political, health, and most importantly technological think tanks.

GEOECONOMIC CHARACTERISTICS OF THE 5G WARFARE

The global economy has been largely devastated by the telecom industry since its beginning. But recently, the new emerging 5G warfare has reached into every corner of it, to the point it started controlling the global economic growth, and ipso-facto became similar to the geoeconomic approaches shaping it.

This warfare, new of its kind as already mentioned, has so many similarities with the soft geoeconomic "trend" of wars recently replacing the military ones on the international scene. Therefore, in the following will be displayed some of the main geoeconomic characteristics of this 5G warfare:

1. The Soft Change

The so-often used concept of soft power, as coined by Joseph Nye in his article in Foreign Policy (Nye, 1990), usually refers to "a nation's ability to co-opt rather than coerce, persuade rather than compel, to set agendas and to attract support". According to him it is composed of three main pillars: a- the appeal of a state's value, b- the legitimacy of its foreign policy, c- the attractiveness of its culture.

Herein, China and the USA are both applying their soft power "values", in order to completely achieve and execute that change in terms of policies and technologies in the "softest" way it could happens on the international scene as battleground.

But due to that, 5G warfare is now called the 5G Cold Technology War, which reflects the fact of its non-military characteristics, turning out to geoeconomics, especially in terms of politico-economic sanctions.

In addition, even the traditional term of soft power is subject to "digital" metamorphosis: it is no more based on political diplomacy, but on cyber – digital one, in the newly created digital sphere. Furthermore, this technological digital soft power is being individually shaped, rather than being exclusively controlled by governments and states.

This transformation is one of the main reasons behind the non-military battle between competitors, because it is relying on individual users, or mass-consumers all over the world.

2. The Fast Adoption, Spread and Expansion

One other main characteristic that distinguishes this warfare, is the fast adoption, spread and expansion of 5G; noting that this digital world is the new global commercial market and has its own economic compact internationally.

According to 5G Americas, "the fifth generation of wireless 5G, powered ahead at four times the speed of subscriber growth as 4G LTE". (Nguyen, 2020) And due to the high effectiveness of 5G on increasing the working productivity, it is becoming extremely essential on our daily lives, starting from social interchange between friends and families, passing by the management of businesses and homes, without forgetting the financial transactions, and even attaining the level of personal healthcare management.

This new network is helping businessmen around the world in building their new strategies and re-configure their own businesses in the post Covid-19 period. In other words, these leaders do not have time to rebuild their damaged businesses and enterprises, therefore, they are recently relying on 5G possibilities, that are the fastest existing solution available for a more efficient and productive future.

Technically, 5G success key is its "faster speed, lower latency and ability to connect vastly higher numbers of devices than previous generations of mobile technology" (Chow, 2021, p. 3).

3. The "Peaceful" Competition (Not So Peaceful Though)

When policymakers of competitive countries around the globe come into adopting 5G and deploying its networks, they stood up facing critical challenges, requiring – or most accurately – obliging them to carefully consider a wide range of political, economic, technical, and strategic respects. Thus, this is shaping the existing race and competition in a more peaceful form, especially that as per Mckinsey Global Institute: "Building a more connected world could create substantial economic value, mostly enabled by advanced connectivity" (MGI, 2020, p.13).

Herein, each pole is seeking to create his own value based on 5G, and that has obliged them to invest all their efforts and time in increasing their own capabilities in this term, which has kept them out of the military fights and clashes.

And as its adoption and rollout is accelerating, the world started witnessing its impact on all terms, mainly politico-economic ones.

Moreover, and because 5G has the potential to enable not only economic growth but enhances the innovation of technologies at mostly all everyday lives' matters… Therefore, China and the US consider it as the key influencing factor of “the great power competition” (Jinghua, 2020).

What differentiates this competition, is that it is happening between companies of these great powers, and not their governments… This makes them “pawns of the geopolitical game” (Jinghua, 2020), which decreases the probability of hard battles, and military wars. Great companies are competing over their positions through their potentials and major capabilities… They undoubtedly are supported by their governments and might transmit data to them. However, it is not in the interest of any government nor its companies to enter into military conflicts due to this competition.

5G PROS AND CONS

Many concerns have been raised due to the spread of 5G, some analyses value its advantages as outweighing its disadvantages, and others see the opposite. This is being a subject of high tension between the two main great competitors, China, and the USA. “Washington has raised concerns that sourcing 5G equipment from Huawei and other Chinese companies will expose a country to national security risks, such as espionage and surveillance. For its part, Beijing has dismissed these concerns as a flagrant attempt to politicize a technological issue” (Tirkey, 2020, p. 4).

This competition was one of the main reasons behind many rumors about 5G pros and cons; especially that the ban over Huawei “also appears to be closely motivated by the US stratagem to prevent China from gaining geopolitical, economic and technological clout by being the first mover of the technology. If Huawei emerges as a leader in 5G technology, China's gains are undeniable and perhaps even inevitable. Beijing may well replace Washington as a leading cyber power, shaping future technological norms for generations to come” (Tirkey, 2020, p. 7).

However, the table 1 will display the major Pros and Cons related to 5G and will be followed by a brief explanation accordingly.

The excitement of developers towards the 5G innovation is understandable to the last extent, but like any network innovation – related to data – many rumbles allege either advantages or disadvantages in terms of this new technology, especially that it was accompanied by the huge pandemic of Covid-19…

However, what will be displayed herein, does not mean that the paper confirms or deny neither the pros nor the cons related to 5G, because it is an extremely critical issue that can be seen from a different perspective. And as every aspect of international relations, 5G should be calibrated objectively. Thus, the paper shows here the fundamental benefits and drawbacks of this innovation as shown by recent studies from different points of view.

Table 1. 5G Pros and Cons.

5G PROS	5G CONS
Increases energy efficiency, eliminates emissions, reduces pollution	Affects health and environment due to exposure to very low-frequency electromagnetic fields
Increases independence and autonomy	Enables control and espionage systems
Improves public safety/emergency response	Enables fast transmissions of military data
Improves health and longer lifespan	Small-cells antennas emit harmful electromagnetic waves and affect health
Increases access to healthcare	Higher radio frequencies (RF) expose individual to health risks
Drives economic growth and sustainability	High costs for 5G rollout and related infrastructure maintenance, in addition to high competition between companies
Shorten commute times	Difficulty of widespread and access to rural areas

5G Pros

Certainly "5G is expected to provide important economic benefits globally". In addition to bringing "substantial networks improvements, including higher connection speeds, mobility, and capacity, as well as low-latency capabilities. In doing so, it enables new use cases and applications that will positively impact different industry sectors" (GSMA, 2018, p. 3).

Therefore, administrations all over the world are taking into consideration the large scope of opportunities that will be afforded by 5G networks in the future. Especially that there are expectations that the implementation of millimeter wave (mmWave) broadbands during 5G applications will have a direct contribution to gross domestic product (GDP) (GSMA, 2018, p. 2).

The previous briefly shows 5G pros in terms of economics and technology, noting that GSMA study concludes, under conservative assumptions, that "by 2034 mmWave spectrum will underlie an increase of $565 billion in global gross domestic product GDP and $152 billion in tax revenue, producing 25% of the value created by 5G" (GSMA, 2018, p**.** 3).

In spite of the claimed bad effects on healthcare – which will be mentioned in the Cons part – 5G is achieving a real technological evolution in the healthcare system, its benefits "will be felt differently by each of the key participants in the healthcare value chain: providers, payers and pharmaceutical companies. But on the whole, 5G networks hold out the promise of major improvements in efficiency and outcomes, positive results that ultimately feed through to patients" (PwC, 2020, p.5).

And with all the innovative applications it provides: Robotics, Internet of things (IoT), and Artificial Intelligence (AI)… the use of 5G will definitely increase in the healthcare system, in a way that will convert it to an ecosystem. This ecosystem "will align with a relatively recent idea known as the 4P medicine: predictive, preventive, personalized and participatory" (PwC, 2020, p.7). In brief, three main advantages of 5G can be owned in the healthcare system if adopted by the healthcare companies: ultra-fast broadband, ultra-low latency, and massive machine connectivity (PwC, 2020, p.11).

Moreover, as mentioned before, and besides gross domestic product GDP, new technology options as faster connection, AI, and IoTs, 5G has a positive impact on peoples' daily lives no matter where they live.

5G Cons

As for the cons, the rumors have been much more numerous, and this is mainly because of the critical competition between the biggest companies and the rival governments supporting them, which has divided the world into a category of people eager to surf on internet at high speeds with all innovative options, and another category scared about the drawbacks of this 5G technology.

And keeping in mind that old phones and devices are not eligible for its application, and that its coverage will not be even and equal to all users all over the world, 5G network has extremely high costs of subscription, where the surfers use more data at higher speed (Gunnarsson, 2020).

"Alongside depleted batteries, users are reporting that cell phones are getting increasingly hot when operating on 5G" (ECN, 2020).

But what is more suspected and worrying are the cybersecurity threats, that can take form in a wide variety of attacks as displayed by the Russian cybersecurity company, Kaspersky (n.d.):

- **Botnet attacks** control a network of connected devices to puppeteer a massive cyberattack.
- **Distributed denial-of-service** (DDoS) overload a network or website to take it offline.
- **Man-in-the-Middle (MiTM) attacks** quietly intercept and change communications between two parties.
- **Location tracking and call interception** can be done if someone knows even a small amount about broadcast paging protocols.

Moreover, health risks have not been studied enough until now, but many claims of health damages were associated to 5G network, especially with Covid-19 pandemic widespread. Although nobody is sure about the nature of side effects that would be caused by 5G waves, but in 2019, and prior to Covid-19 outbreak, Dariusz Leszczynski, an expert in molecular biology and Adjunct Professor at the University of Helsinki, Finland, told Euronews that "the assurances of safety concerning 5G-emitted radiation are based solely on the assumption that low amounts of radiation are safe, not on biomedical research" (Beswick & Fischer, 2019).

Furthermore, and since its outbreak, Covid-19 was linked to 5G in most of the analyses, claiming that 5G frequencies could exacerbate its spread and suppress the immune system. (Kennedy, 2020).

However, and according to the last statements given by the World Health Organization WHO "and after much research performed, no adverse health effect has been causally linked with exposure to wireless technologies. (…) Tissue heating is the main mechanism of interaction between radiofrequency fields and the human body. Radiofrequency exposure levels from current technologies result in negligible temperature rise in the human body. As the frequency increases, there is less penetration into the body tissues and absorption of the energy becomes more confined to the surface of the body (skin and eye). Provided that the overall exposure remains below international guidelines, no consequences for public health are anticipated" (WHO, 2020).

CYBER-DIPLOMACY METHODS

Thinking about all these conspiracies linked to 5G, especially in terms of cybersecurity – whether they were true or baseless – brings into mind the cyberspaces of competing states in addition to their ways in putting cyber norms and facing cyber threats. Noting that at the London Summit in 2019, the 29

members of NATO agreed to “guarantee the security of our communications, including 5G” (Dinucci, 2019).

Thus, cybersecurity is a matter of high interest for all. And in order for them to protect their interests and to overcome surprising challenges, they referred to a distinct type of interaction: Cyber-diplomacy!

It is necessary to understand how the United States and China are figuring out a way to manage their concerns of security during the peak of their competition over technology, especially that The NATO Cooperative Cyber Defense Centre of Excellence quotes ancient Chinese philosopher Sun Tzu's “The Art of War,” to summarize its opinion on Huawei's and China's real intention: “The supreme art of war is to subdue the enemy without fighting” (Kaska et al., 2019, p. 4).

Cyber-diplomacy is the international emerging practice used by the US and China during the current escalation of the national security debate, noting that “a potential threat anywhere in the network will be a threat to the whole network” (Reichert, 2018). And none of these competitors accepts to be threatened in any way, so they used cyber-diplomacy even before the release of 5G to overcome the expected hurdles.

“Cyber-diplomacy can be defined as diplomacy in the cyber domain or, in other words, the use of diplomatic resources and the performance of diplomatic functions to secure national interests with regard to the cyberspace” (Barrinha & Renard, 2017, p.355). Practically, it is a natural response to the increasing relevance to cyberspace globally, especially after being threatened recently, and “through cyber-diplomacy states collaborate to respond to and addresses the cyber dimensions of international conflicts, crime, and information security” (Khabbaz, 21, p. 1).

USA’s Cyber-Diplomacy

USA makes a great example for states relying on cyber-diplomacy for their cyberspace security: it has cooperated with other European states since the pre-release of 5G and developed its “National Strategy to Secure 5G” where it called for adopting the “5G security principles” outlined in the 2019 Prague Security Conference: Prague Proposals”. (Khabbaz, 21, p. 1).

So, in March 2020, while Donald Trump was still president, the US National Strategy to Secure 5G was released, and it included 4 lines of efforts: 1- Facilitate Domestic 5G Rollout; 2- Assess Risks & Identify Core Security Principles of 5G Infrastructure; 3- Address Risks to United States Economic and National Security During Deployment of 5G Infrastructure Worldwide; 4- Promote Responsible Global Development and Deployment of 5G. (White House, 2020).

Much more, one of the main activities of the National Strategy was a plan for Diplomatic Engagement, where they have included diplomatic activities with partner countries and allies, based on “diplomatic engagement to share information and findings on 5th and future generations wireless communications systems and infrastructure equipment standards to promote maximum interoperability, competitiveness, openness, and secure platforms” (White House, 2020). Eight main elements formed the key tactics of this plan:

1. Raise Awareness Among Allies and Partners on Security Risks.
2. Encourage Allies and Partners to Take Concrete Actions to Protect their 5G Networks.
3. Encourage use of Trusted 5G Vendors.
4. Partner with Like-Minded Countries.

5. Public Diplomacy.
6. Promote the Prague Proposals.
7. Multilateral Engagement.
8. Encourage Allies and Partners to Require a “5G Clean Path” for Overseas Facilities.

This diplomatic plan was thoroughly based on security measures concerning the cyberspace of US and its allies, and on US national interests. It is quite different from a normal diplomacy plan because it addresses technical matters rather than political ones, even if in some areas they are convergent and complement themselves – especially of the politico-economic impact of 5G implementation – but it is still implemented by diplomats of all involved governments. 5G was referred here to technical matters, because “cyber issues were treated f irst as purely technical issues, then as external aspects of domestic policies, before they came recognized as a major foreign policy topic” (Barrinha & Renard, 2017, p. 358).

Therefore, coordinating US’s diplomatic engagement on cyber issues related to 5G is a remarkable act, especially that it has attained China since 2015 with a bilateral Cybersecurity agreement concerning economic espionage (Brown & Yung, 2017). After having accusing Huawei of abusing the Cybersecurity Law and prohibiting several government agencies since 2012 “on the grounds of national security risk, from acquiring products from Huawei and ZTE” (Rugge, 2020, p. 4).

China’s Cyber-Diplomacy

China’s approach to cyber-diplomacy differs from that of the USA, for it is being more offensive than defensive. It is driven by the “overarching objective of become a cyber superpower in the economic, normative, military and commercial realms – one that harnesses the power of digital technologies and innovation to achieve global technological leadership and modernize economic development” (Bozhkov, 2020).

However, and putting the objective aside, China’s cyber-diplomacy is based on its promotion for Cyber-sovereignty “as an organizing principle of internet governance, in direct opposition to US for a global, open and secure internet”. And unlike the USA who exclusively implements cyber-diplomacy with its allies and partners countries, China went to the UN in 2017, and called for a “multilateral approach to governing cyberspace, with the United Nations taking a leading role in building international consensus on rules” (Segal, 2017).

Also, and after the opposition of US and its allies to China’s idea of state sovereignty in cyberspace, China’s representative noted in his opening statement at the Open-Ended Working Group (OEWG) meeting in September 2019 that it was “widely endorsed by the international community that the principle of cyber sovereignty applies in cyberspace” (Segal, 2020, p. 3).

From another side, we can say that China has not only referred to the international community through the UN only, but has also worked on its cyber-diplomacy strategy through its allies, mainly Russia, but in a method different than the American one. Beijing worked on Moscow to become a promoter for its “ideas” and opinions at the UN too, in a way that convenes its national interests; this is why Russia worked on promoting the Cybercrime treaty at the UN (Segal, 2020, p. 4).

Furthermore, China was not content with cyber diplomacy only, but combined commercial diplomacy to it. And this appears clearly in the “digital silk road” of the Belt and Road initiative, where Chinese

companies invest in “cross-board optical cables and other communications trunk line networks, transcontinental submarine optical cable projects” (Segal, 2020, p. 4).

By this, one can understand that Chinese companies are searching for new markets and customers for their 5G items and products, while the Chinese government provides support to the BRI countries in terms of economy, strategy and sometimes politics. And according to the Financial Times magazine, “China’s Export-Import Bank financed 85% of the China-Pakistan Fiber-Optic Project, for example and loaned to Nigeria the full cost of a Huawei-built 5G network” (Weinland, 2019).

This paper cannot forget to mention neither the “smart cities” that China is working on building across many states, and through which it competes with suppliers from Europe and US, nor the leadership of Chinese firms in terms of AI surveillance technology used for public security (Segal, 2020, p. 5).

For concluding, when a superpower releases a national strategy to combat 5G threats, and another one promotes 5G rollout and requires surveillance, this means that the challenge is real, and 5G is a new international topic over which superpowers are typically competing.

DISCUSSION

Ex-President of the United States, Donald Trump has expressed many times his worries towards the escalating economic capabilities of China… And herein, Biden came to thoughts… But as Michael Mcloughlin (2020) stated: “Biden's victory doesn't look like it will mean much change on the US side”. But the world’s attention is now focused on the way in which Joe Biden will manage this “revolution”.

As far as tech industry is concerned, Google has finished its relations with Huawei after Trump’s veto; and so, this has left Huawei’s devices without the mobile services and apps customized by Google. However, due to the huge difference in costs of other brands devices compared to Huawei’s ones, market sales have suffered, and phone shipment orders have globally decreased by 23% in the third quarter of 2020, according to Canalys, and by 24% according to Counterpoint Research (Mcloughlin, 2020). That is why, dealers are obliged to neglect the decision of banning Huawei’s devices in their countries, for their own benefit, and that of the customers.

In addition: “Now, according to Financial Times, tough U.S. sanctions this year against Huawei could be less threatening to its overall business than previously thought. And according to analysts, the company’s important smartphone arm might have a chance to recover. Less threatening to Huawei means more threatening to Google and its lock on the worldwide Android ecosystem” (Doffman, 2020).

In terms of healthcare, recent observations have raised concerns regarding the 5G effect on the human health, especially after the mystery of Covid-19. Many debates are arguing on either it is really dangerous or does not have any negative repercussion on human health. This study does not settle the debate but represents a comparison between the pros and cons of 5G evolution on the healthcare system, in this period.

Thus, 5G has succeeded in creating a vision to turning out healthcare system to SMART one. And according to an article shared by Oxford Academic, one month before the first Chinese infection of Covid-19 has been identified: “5G will reconstruct the healthcare system by intelligently improving the quality of medical service, balancing the distribution of medical resources between urban and rural areas, and reducing the burden of healthcare costs” (Li, 2019).

CONCLUSION

In summary, in terms of economics, this paper shows that the gradual rise of 5G and its innovations, coincided with slowing down productivity and the Gross Domestic Product growth in developed countries. Indeed, one analysis mentions that “the US economy loses $ 1tn each year due to too much information and interruption” (Konzept, 2019, p. 7). So, guess its losses during 5G phase!

Technologically speaking, 5G is considered as a revolution of a new kind: “this new technology will be the backbone that enables advances such as smart cities, driverless cars, remote controlled operating theatres, automated farms and more besides” (Cooper, 2018).

In terms of healthcare, this study agrees that “there is no concrete evidence of health damage due to 5G electromagnetic waves, although it is not scientific and illogical to prove the non-existence of negative effects” as Prof Dr. Alexander Lerchl told Euronews (Beswick & Fischer, 2019).

But this study realizes that, the fear of 5G network, is more linked to politico-economic balances between Nations than related to health – Notice the USA has taken a banning decision against Huawei, but never against its main alternatives Nokia or Ericsson. “The risk is not the same with European manufacturers as with non-European ones” said Macron's team (Mcloughlin, 2020).

However, and regardless of superpowers interests, it is better for every individual to be cautious from their blind race for influence.

KEYWORDS: 5G, Cyber-diplomacy, Digital Sovereignty, Geoeconomics, Techno-war.

REFERENCES

Barrinha, A., & Renard, T. (2017, December 28). *Cyber-diplomacy: the making of an international society in the digital age.* Global Affairs, Routledge.

Beswick, E., & Fischer, L. (2019, March 26). *What are the health risks associated with a 5G network?* Euronews. https://www.euronews.com/2019/03/26/what-are-the-health-risks-associated-with-a-5g-network

Bozhkov, N. (2020, March 9). *China’s Cyber-Diplomacy: A Primer.* EU Cyber Direct: https://eucyberdirect.eu/content_research/chinas-cyber-diplomacy-a-primer/

Bremmer, I. (2019, May 24). *The Quick Read About… the 5G War Is Upon Us.* Time Magazine. https://time.com/5595161/the-quick-read-about-the-5g-war-is-upon-us/

Brown, G., & Yung, C. D. (2017, January 21). Evaluating the US-China cybersecurity agreement, Part 3. The Diplomat. https://thediplomat.com/2017/01/evaluating-the-us-china-cybersecurity-agreement-part-3/

Chow, W. (2021). Executive Summary. *The Global Economic Impact of 5G.* PWC, China.

Cooper, M. (2018). 5G Technology Briefing. *ITNOW*, *60*(3), 16-20.

Dinnucci, M. (2019). *The Hidden Military Use of 5G Technology,* Telesur. https://www.telesurenglish.net/opinion/The-Hidden-Military-Use-of-5G-Technology-20191221-0006.html

Doffman, Z. (2020, November 1). *Huawei’s New ‘Fight’ With Google To Beat Android Is Suddenly More Threatening*. Forbes. https://www.forbes.com/sites/zakdoffman/2020/11/01/huawei-wants-to-beat-google-android-apple-iphone-and-samsung-galaxy-after-trump-ban-update/?sh=4ae79900e39e

ECN (2020, July 28). *What Are the Disadvantages of 5G? 6 Disadvantages of 5G.* https://www.ecn.co.za/what-are-the-disadvantages-of-5g/

GSMA (2018, December). *Study on Socio-Economic Benefits of 5G Services Provided in mmWave Brands.* UK

Gunnarsson, E. (2020, August 31). *5G – 5 Drawbacks You May Not Have Thought Of.* Soluno https://www.soluno.se/en/5g-5-drawbacks/

Jinghua, L. (2020, September). *The Race of Chinese Companies in the 5G Competition.* Italian Institute for International Political Studies. https://www.ispionline.it/en/pubblicazione/race-chinese-companies-5g-competition-27511

Kaska, K., Beckvard, H., & Minarik, T. (2019). *Huawei, 5G and China as a Security Threat.* NATO Cooperative Cyber Defence Center of Excellence CCDCOE. https://ccdcoe.org/uploads/2019/03/CCDCOE-Huawei-2019-03-28-FINAL.pdf

Kaspersky. (n.d.). *Is 5G Technology Dangerous? - Pros and Cons of 5G Network*. https://me-en.kaspersky.com/resource-center/threats/5g-pros-and-cons

Kennedy, R. (2020, April 5). *People are saying coronavirus is a cover up for 5G – here is why it's not.* Euronews. https://www.euronews.com/2020/04/04/is-coronavirus-a-cover-up-for-deadly-effects-of-5g-technology-euronews-answers

Khabbaz, D. (2021). *Cyber-diplomacy: Benefits, Developments and Challenges,* Yale Law School. https://static1.squarespace.com/static/58011083197aea95712e1bf4/t/5ed5230bb568f07c9647c189/1591026443959/Cyber+Diplomacy+Benefits%2C+Developments%2C+and+Challenges.pdf

Konzept. (2019). *How will 5G change your life* (16). Deutsche Bank Research.

Li, D. (2019). 5G and intelligence medicine—how the next generation of wireless technology will reconstruct healthcare? *Precision Clinical Medicine*, *2*(4), 205-208.

McKinsey Global Institute (MGI), (2020, February). *Connected World: An Evolution in Connectivity Beyond the 5G Revolution,*

McLoughlin, M. (2020, November 16). *Trump pierde, ¿Huawei gana? Qué ocurrirá ahora con el despliegue mundial del 5G*. El Confidencial. https://www.elconfidencial.com/tecnologia/2020-11-16/gran-guerra-5g-trump-huawei-china-biden_2825419/

Nguyen, V. (2020), *5G is the Fastest Growing Mobile Technology in History.* 5G Americas, in Globe News Wire. www.shorturl.at/gEMPQ

Nye, J. S. (1990). *Soft Power.* Foreign Policy. https://www.wilsoncenter.org/sites/default/files/media/documents/page/joseph_nye_soft_power_journal.pdf

Pant, H. (2020), Preface of *The 5G Dilemma: Mapping Responses Across the World.* Observer Research Foundation, India.

PwC (2020). *5G in Healthcare – How the new wireless standard can connect a post-covid healthcare ecosystem.* China. https://www.pwc.com/gx/en/industries/tmt/5g/pwc-5g-in-healthcare.pdf

Rathbun, B. C. (2008). Interviewing and qualitative field methods: pragmatism and practicalities. In *the Oxford handbook of political methodology*.

Reichert, C. (2018, November 5). *Huawei Denies Foreign Network Hack Reports.* ZDNet: https://www.zdnet.com/article/huawei-denies-foreign-network-hack-reports/

Rugge, F. (2020, September). *5G in a Contested Domain.* In ISPI Dossier: The Geopolitics of 5G, Italian Institute for International Political Studies.

Segal, A. (2018, September). *When China Rules the Web". Foreign Affairs* https://www.foreignaffairs.com/articles/china/2018-08-13/when-china-rules-web

Segal, A. (2020, March). *China's Alternative Cyber Governance Regime,* Council on Foreign Relations – CFR.

Thales. (2020, December 3). *Introducing 5G technology and networks (speed, use cases and rollout)*. Thales Group. https://www.thalesgroup.com/en/markets/digital-identity-and-security/mobile/inspired/5G

Tirkey, A. (2020) *The 5G Dilemma: Mapping Responses Across the World.* Observer Research Foundation, India.

U.S.-China ESRC: Economic and Security Review Commission. (2020). *2020 Annual Report to Congress, 116th Congress, Second Session*. https://www.uscc.gov/

Weinland, D. (2019). *China State Banks Pull Back from Risky Overseas Projects.* The Financial Times: https://www.ft.com/content/273c324c-55ec-11e9-a3db-1fe89bedc16e

White House. (2020, March). *National Strategy to Secure 5G of the United States of America.* https://www.ntia.gov/files/ntia/publications/2021-1-12_115445_national_strategy_to_secure_5g_implementation_plan_and_annexes_a_f_final.pdf

WHO – World Health Organization (2020, February 27). *Radiation: 5G Mobile Network and Health.* https://www.who.int/news-room/q-a-detail/radiation-5g-mobile-networks-and-health

SECURITY UPDATES FOR ENHANCEMENT ON TRUST AND CONFIDENCE IN E-LEARNING SYSTEMS

Nuno S. Silva, Isabel Alvarez

Universidade Lusiada de Lisboa (Portugal), COMEGI / ISTEC (Portugal)

nsas@lis.ulusiada.pt; alvarez@edu.ulusiada.pt

ABSTRACT

The COVID-19 pandemic situation emphasized the need for a strategic response to change educational context, with Governments and educational institutions intending to use this vision and e-learning. This environment is noticeable at three main levels: micro (e.g. the relationship between lecturers and students); meso (e.g. the existing electronic universities' projects); and macro (worldwide governmental actions). Moreover, e-learning, with the utilisation of ICT, causes universities to think globally (competitiveness) and internationally (collaboration). However, to presume that technology by itself entails education is unrealistic and condemns any e-learning paradigm, because novel technologies impose substantial security issues. This work intends to plan a risk mitigation regarding e-learning security implementation. The confidence in the availability, and non-repudiation, should be combined with the aim of reaching information security requirements for e-learning systems as a precondition for enhanced user acceptance. Trust is also essential to get the better of privacy concerns.

INTRODUCTION

In e-learning, attention should be paid to security elements such as availability, integrity and confidentiality to avoid any security breaches that may harm educational institutions (Okafor, Oparah & Okwudili, 2018). The credibility of online learning and the privacy of students and staff need to the considered. Any e-learning system is supported by the Internet which is insecure and which opens a door to any software attack. However not much has been done to avoid this situation. To guarantee a safer e-learning environment and avoid these threats, security should be a priority.

E-learning promotes the existence of a strategic response to a novel educational context, which is emphasised by the COVID-19 pandemic (Almaiah, Al-Khasawneh, and Althunibat, 2020). Governments and educational institutions intend to use this vision. The use of Information and Communication Technologies (ICT) mean large changes to the way lecturers, students and universities use to work. However, to presume that technology by itself entails education is unrealistic and condemns any e-learning paradigm, as instructional quality, privacy, and mobility need substantial security issues with the use of technology. Moreover, E-learning causes universities to think globally (competitiveness) and internationally (collaboration), although these distributed knowledge networks may cause several ethical and cultural dilemmas.

This work suggests a risk mitigation plan regarding e-learning security implementation; we propose the exploration of four levels: Technological Infrastructures and Services, Knowledge/Content Management, Computer Mediated Communication, and Value-added.

To mitigate web attacks on e-learning platforms, Wani & Khan (2016) refer the critical security issues that an e-learning platform must address and a methodology of development that should incorporate security components at the design phase to avoid these threats.

DIMENSIONS OF E-LEARNING

Overview

The e-learning literature is vast. Several authors refer that this form of learning currently depends on networks and computers but will likely evolve into systems consisting of a variety of channels. It is essential to understand the global perspectives of e-learning intervention, as, for example, it refers to the complex connections between strategies, design, and technologies, which encompass the following components of policymaking: strategic planning and vision; curriculum and content; use of the internet and acceptable use policies; quality assurance and accreditation; conductivity, infrastructure, and networks; professional development; intellectual property and copyright; cost, finance, and partnerships; factors such as leadership, culture, structure, design and technology, as well as delivery management. In spite of this, it is suggested that four levels are explored: Technological Infrastructures and Services, Knowledge/Content Management, Computer Mediated Communication, and Value-added.

Technical Infrastructure and services

As for technological infrastructure and services, the e-learning implementation at university settings is a complex task, which starts with a strategy for developing the basic technical infrastructure. According to Blinco, Mason, McLean & Wilson (2004, p. 2), this “infrastructure often describes a bottom layer of an architectural description or diagram, indicating network hardware components, communication processes, services and protocols”. Throughout this assumption it is vital to shed some light over what are the issues in a "bottom layer". First of all, it needs a comprehensive functional and technical analysis to determine how technology should be applied, where equity of access is an important factor, and what security levels are to be implemented (e.g. firewalls).

Although the internet is the basic network infrastructure for e-learning it is, however, necessary to consider components at local networks, as well as personal tools and equipment that make learning activity possible. Management systems assume particular importance on this assumption as the e-learning evolution comprise four general categories of technological systems: Learning Management Systems, Managed Learning Environment, Learning Content Management Systems and Virtual Learning Environments. Furthermore, several other dilemmas may emerge asking for a global open source solution (like, for instance, Google).

Knowledge/content management

In what concerns knowledge/content management, e-learning also includes content over technology, or, following Hartley (2014) educational content is more important than technologies. This assumption leads us to explore content related issues in e-learning implementation, and thereby the need for a new conception of security understanding.

It is discussed and argued that a clear perception of the boundaries of knowledge versus content is required, since these two concepts also influence pedagogical strategies in e-learning practices. If on

one side knowledge is dependent on conceptual skills and cognitive abilities, on the other content refers to the encoded "unprocessed material" which achieves the objectives that the content creator has set for it.

In traditional learning approaches (behaviourism, cognitivism and constructivism), and aiming that learners master knowledge through drill and practice, content is categorized according to its encoded meaning; it should be measured by the fulfilment of the end goals as being highly interactive and add value. In an e-learning implementation both these two statements, due to the interactivity potential of computer-mediated communications, are subject to transformation in what concerns the knowledge/content creation.

Computer-mediated communication

In computer-mediated communication, the interactivity is a key characteristic of e-learning within the communication processes to be analysed. Computer Mediated Communication (CMC) should be explored in regard to one additional layer to understand the e-learning implementation. According to Zhang (2004), CMC transforms classrooms to make learning a more interactive, diverse and enjoyable experience. This can be through online interactive classrooms, interactive group discussions and tutor/student sessions, or empowering students/teachers' interactions by designing more flexible and intuitive interfaces. In this educational paradigm, learning "with" interactive technologies establishes a certain intellectual affiliation between students and technologies. Instead of using technologies to guide students through prearranged interactions, students may use technologies that function as "the mindful engagement of students". The social presence created in an online community was a strong predictor of satisfaction in CMC, building learning communities. However, how can CMC support satisfactory socio-emotional and relational communication compared with face-to face communication. How can this evolution of learning through technologies replace teaching as essentially a human relation, face to face relationships, and social isolation in front of a computer screen. In spite of robot deployed as classroom teachers, there is a risk that children would lose emotional security (Sharkey, 2016).

We need to implement a balanced approach to avoid ignoring technology tools or fixating too much on technology for e-learning. The videoconferencing is considered as the CMC tool that is the closest to face-to-face communication, enabling high levels of interaction and facilitating personal feelings (e.g. social presence and perceived privacy), while security breaches are usually out of risk analysis and its related streaming media services, like the recorded classroom lectures or video-broadcast seem to hold an important value.

However, following May et al. (2012) using a CMC tool alone, participants claim that their activities are not fully controlled as when learning in a traditional face-to-face situation. In what concerns tracking technologies, if users are informed in anticipation of any tracking process being used in the platform that they are accessing, having the possibility to give their approval for this tracking process to take place, then they should not consider these technologies as a threat. It is assumed that in a collaborative learning situation where intense interactions and exchanges of both personal and collaborative data take place, there should be a specifically designed environment to guarantee the protection of the learner's privacy.

Value added

Value added emerged as an important approach to e-learning (for instance related to time investment costs, content valuable functionality, and use of streaming media services). So, in this scenario, it is possible to understand that e-learning adds value to the learning experience, but it is not clear what issues mean “value of e-learning” (financial notion on the measure of benefit), or “e-learning values” (ethically and culturally sensitive to meanings that may vary according to context). Therefore, it is important to explore if it justifies a “top layer” to understand the e-learning implementation.

In late 2019 a novel worldwide health situation unexpectedly emerged: the sarscov2, a new type of coronavirus was identified (WHO, 2020). Due to the pandemic restrictions, e-learning was enforced as the only way to proceed with learning. Based on this experience, most people are looking at this opportunity to introduce and implement innovative alterations in the education system (Bird & Bhardwaj, 2020). Researchers should take this opportunity to understand and explore the crisis implications in terms of learning, evaluate the impacts and the financial consequences of this pandemic and develop models of policy interventions (Brammer & Clark, 2020). In the outcome, it can be learned from this pandemic experience, that they could adjust to different environments and situations, facing and solving new difficulties and conclude that positive perspectives arose from the quarantines.

From the teacher’s perspective, the increasing use of digital platforms like the interactive meeting tools – Zoom, Google Meet, Microsoft Teams or Cisco Webex – was seen as the best way to continue with learning. After this very interesting pandemic experience, some advantages can be seen in the use of online learning, like the flexibility of time and place, inducing to reflexive innovations in the practices and schedules of academic structure (Brammer & Clark, 2020). However, the impact of COVID-19 in education cannot yet be fully understood (eLearning Inside, 2021; Zhao & Watterston, 2021). Moreover, in what concerns security breaches of this impact needs a research focus that is not yet developed.

Xie (2020) refers some of the threats of online education: network instability and technological constraints; lack of a sense of belonging and connectedness; presence of distractions. It can be considered that this pandemic scenario provided a chance to improve their information literacy and security awareness. Having to learn how to use new tools and software that they were not acquainted before, was indeed a “learning opportunity” leading to thoughtful decisions in the future of e-learning.

E-learning and Cloud computing

From an educational point of view, and acknowledging that cloud computing might provide unlimited resources for storage capacity and data processing, universities will become a key element in future security systems. A variety of services through pay per use and fee-based infrastructure with value added infrastructure will be offered to the users of cloud computing (Arora & Sharma, 2013), as this innovative technology delivers computing resources through globalised circulation networks.

The introduction of cloud computing in the educational systems acknowledges the purpose of increasing scalability, flexibility and availability at the application level (Popel & Shyshkina, 2019).

In addition, cloud-based LCMS (Learning Content management Systems) centralizes content management, while providing in-depth performance metrics, ensuring resolution in time to execute developments, support, updates and fixes.

Cloud computing increasing interest has implications for security, privacy, and trust. Compliance with European GDPR requirements is a critical requirement. COVID-19 caused the criticality of technology to increase the use of platforms such as VPN, video conferencing tools, and home computer equipment. This implies that the SLAs associated with these environments must be improved. Even so, governance of emerging technologies is critical to undertake advanced measures to protect the most security-sensitive information stored (Rawtani, 2012). The confidence in the availability, and non-repudiation, should be combined with the aim of reaching information security requirements for e-learning systems as a precondition for enhanced user acceptance (Moneo, et al. 2016; Weippl, 2005). In addition, the diversity of mobile devices and their security protection measures are varied in accordance with the operating system (May Iksal & Usener, 2017), and biometric web authentication can be useful for proper identification of learners (Goyal & Krishnamurthi, 2019).

In spite of the foreseen important short and long-term benefits that cloud computing will bring to a scholarly environment, being an emergent technology, serious perils and ethical challenges may occur and need to be considered. The increasing interest by policymakers and regulatory authorities in cloud computing is due to the possible implications in security, privacy and trust (Rand Europe, 2011). Although privacy may yet be in opposite laws and debates (EDRI, 2021; Salter, 2019), while not only the legality but also the morality is important in cyber security (Hamburg & Grosch, 2017).

Indeed, the governance of emerging technologies, whose menace can merely be lessened, is critical to undertake advanced measures to protect the most security-sensitive information stored (Rawtani, 2012). A responsible management of personal data becomes a key issue to ensure trust in cloud-based services adoption and encouraging users to explore them (Pearson, 2013).

Data security covers main areas like encryption, and password security. It should be considered to the appropriate balance between law practices and protecting confidential data in cloud-based storage.

It is relevant to report problems for instance in SSO (Single-Sign-On) synchronisation, content synchronisation, trust of passwords (external security), and the unexpected software updates; some novel functionalities of cloud computing desperate users due to the lack of usability!

Furthermore it can be merged with high compromise of top management in security strategies which made part of ISO/IEC 31000 and ISO/IEC 27005 risk assessment standards.

Trust and confidence

Concerning trust management, various issues should be taken in consideration like privacy of users (learners and teachers) including issues related to grading, competency, and personal information (Wani & Khan. (2016). Itani, et al. (2014) for instance suggest that the minimum service metrics (reliability, availability, performance and security) for cloud service provider based on SLAs (Service Level Agreement) which, if fails, will affect reputation and trust.

May & George (2011) suggest that trust is part of the solution. Trust is confidence on someone's competence and commitment to achieve a goal and crucial to build and assist relevant interactions in learners collaboration.

Trust is also essential to get the better of privacy concerns on using technological solutions and not only to keep users safe from any threats. In practice, privacy and trust are circularly related, as privacy is a natural concern at the same time that trust is an essential factor in the learning environment. In fact, in a closed learning environment, where all learning services are provided internally (e.g. from a university or a trusted source) students can have higher confidence that their personal data and

learning tasks such as working collaboratively will be treated properly On the other hand, in an open learning environment with unknown providers such as private or external learning service providers, privacy concerns are higher and the trust level of learners will be influenced by the perceived privacy offered by those providers; these privacy issues concern learning technology providers, learning service and content providers and the participants as well, with the crucial tasks for learning service and content providers being to secure learning environment and storage of learner data. As for the participants, they are mainly concerned with the trust assessment of the learning environments they are using, and with the protection of their sensitive personal data; also understanding the security issues in the learning situations helps them to avoid security threats as well as to improve protection of both themselves and their learning environments. Security and privacy levels differ according to the various learning environments and depend on the types of learning activities being done by the participants. From May (2011), the following issues concerning learning technology: personal data protection, anonymous use of learning services, address and location privacy, single sign-on, seamless access to learning resources, authenticity of learning resources (Las), digital rights management, legislation and awareness raising require different protection provisions.

Presently students have an increasing understanding and knowledge of information systems (IS) and information technology (IT) issues. To build digital trust and to assist the students’ needs all the learning strategies devised by course providers must be linked with IS/IT strategies whether now or in the future (Bandara, Ioras, & Maher, 2014). In terms of usability, security and protection of their personal information, digital natives and immigrants will share expectations of their e-learning systems; this could refer, for instance, the inclusion of students’ details associated with payments for course fees and other products, done in a secure way. The important intellectual property connected with research and other academic material existent in the UK Universities could attract cyber-criminals and researchers will expect that their important work and sensitive information is properly and securely stored, with no risk of theft or misuse. Institutions should perform a cyber security risk assessment and determine best arrangements for technology, people and processes.

To plan a risk mitigation regarding e-learning security implementation

To plan a risk mitigation regarding e-learning security implementation, there are several issues to be considered: Internet bandwidth must be recognised as fundamental, alongside issues like speed, accessibility, cost and reliability; wireless networks and mobile computing for students is a key benefit because it avoids the need for physical presence; videoconference implementation enabling online-only teaching (e.g. Microsoft Teams, Zoom, Google Meet or Cisco Webex); and the cyberthreats of online learning platforms, as for example Moodle e-learning environment added value related to access, privacy and security, since lecturers can make content available only for students who must take the inherent access rights for course units. Concerning ethical assessment of the idea of robot teachers, Sharkey (2016) report the relevance of issues including privacy, control, and accountability.

It is also typical to report problems at meso level, for instance in SSO (Single-Sign-On) synchronization, content synchronisation, trust of passwords (PKI and external security) (Miguel, Caballé, & Xhafa, 2017), and the unexpected software updates; also a self-service integrated system allows printing, photocopying and scanning of content, where security is based on password request (security breaches in networked printing systems). Another aspect of security is the existing backup policy. Moreover, it is important to mitigate the risk rating of all external assets, such as web applications, IP addresses, and marketing sites. For example, a Moodle penetration testing when performed can reveal important issues for risk monitorization. This study intends to merge a new security sensitivity

metric for such variables. It includes security updates and best practices that effectively minimize risks related to using cloud computing.

Schinagl, Schoon, & Paans (2015) developed a measurement method to assess the effectiveness of the protection provided by a SOC (Security Operations Centre), which is responsible for the activities related to security monitoring as well as to address situations that jeopardize the confidentiality, integrity and availability of technological services and electronic information that the university administers. The SOC staff to cover the following functions: Manager - the group leader that can assume any role while overseeing general security systems and procedures; Analyst - Analysts compile and analyse data, either from a period of time or after a breach; Investigator - Once a violation occurs, the investigator finds out what happened and why, working closely with the responder; Responder - Tasks that come with responding to a security breach; Auditor - Current and future legislation comes with compliance mandates. This feature can keep up to date these requirements and ensures that the organization meets them.

According to NIST (2012), a computer security incident is the imminent violation or threat to an information security policy violation (acceptable use policies or standard security practices).

Given such a reality, it is important to refer to two levels of arguments concerning e-learning security updates:

- as a strategy - organisational change addressing security issues;
- as a tool - socio-technical dimension of security breaches (password security level and data protection).

Therefore, in order to minimise potential failures, it is crucial to involve all stakeholders to have security awareness and security sensitivity as a prerequisite for trust and confidence in national and international successful e-learning implementation.

The mobile services available anywhere around the world, has been impacted by the continuous growth in the numbers of smart devices and related connectivity loads. Authentication is in fact, in such a connected globe, in the first place, the enabler keeping the transmitted data secure (Alomar, Alsaleh, & Alarifi, 2017).

Actually, there are three factor groups available to connect an individual with the established credentials (Harini, & Padmanabhan, 2013):

1. Knowledge factor—something that is only knwn to the user, such as a password or, simply, a “secret”;
2. Ownership factor—something the user has, such as cards, smartphones, or other tokens;
3. Biometric factor—something the user is: it could be biometric data or a behaviour pattern.

As suggested by Mohsin et al. (2017), to provide a better level of security and to ease continuous protection of computing devices together with other critical services from unauthorized access by using more than two categories of credentials, the Multi-Factor Authentication (MFA) is proposed.

Confidentiality, Integrity, and Availability, are the basis of all security programs and security in computing concentrates on tools, processes and methods to design, develop and implement reliable systems (Asmaa, & Najib, 2016). Examining Moodle, one of the most used and popular open-source

e-learning systems - a system needs to implement security services such as authentication, encryption, access control, managing users and their permissions. However, several vulnerabilities are historically reported (CVE, 2021), with statistics put in the top for Moodle both "Gain information" (allows remote attackers to obtain sensitive information), and XSS (an attacker to compromise the interactions that users have).

To conclude, a secure e-learning platform should include all the types of security producing transparent processes both to the teacher and student. It is assumed that in the future the concept of m-learning will come in new electronic learning features but, in parallel, new risks will also occur.

Information security and privacy such as confidentiality, integrity, authentication, and non-repudiation, is no longer only a highly desired feature but it is now an essential legal required condition of any information system, with e-learning being not an exception (Wani, & Khan, 2016). Any e-learning platform is web based and therefore is prone to diverse attacks, such as brute force attacks, XSS (or Cross Side Scripting), direct SQL code injection, remote SQL injection using a virus/trojan file, SQL injection in the site address CURL SQL injection), web indexing, session predictions, password cracking, etc. At present, not many e-learning systems have proper security design and privacy characteristics integrated into the e-learning development and implementation process. Consequently, such systems can cause serious issues in maintenance of learning objects, authentication in student/teacher registration, scheduling of events, protection of user profiles, conduct of remote assessments and examinations, and certification award. Vulnerabilities can be accidentally and intentionally introduced throughout the software development life cycle during requirements definition, design, implementation, deployment and maintenance and e-learning systems are vulnerable to a number of threats: serious security threats include software attacks (viruses, worms, macros, denial of service), espionage, acts of theft (illegal equipment or information) and intellectual property (piracy, copyright, infringement) (Bandara, Ioras, & Maher, 2014).

The learner data is normally used to enhance the learner's security position by continuous delivery of important information and accommodating security mechanisms (Milošević, & Milošević, 2016). From the list of presented protection measures, the most urgent ones are those related to potential security and privacy threats and protection of personal data, while the least urgent (but still relevant) is the anonymous use of learning resources. An e-learning system being a web-based system and must be protected from any computer threats to ensure the tranquillity of the users when using it, with or without online webcams.

This study has also investigated the security benefits and increased threats of shifting from traditional monolithic system to modern e-learning ecosystem or cloud-based system, also examining loss or theft of mobile device, unauthorised access, attack on m-learning system and denial of service.

Finally, Universities have to carry out actions and create resources that promote information security at micro level (MetaRed, 2021), and for that, the recommended example is the Cybersecurity Awareness Kit developed for Ibero-American Universities as a collaborative project with Spanish Cybersecurity Institute (INCIBE).

Therefore, there are dilemmas to conceive e-learning goal from the macro level, namely if it enlarges the scope of traditional university. For the purpose of innovation and change in the university role it may enhance the traditional forms of university teaching and administration (hybridisation), but the user's perspective early in the implementation process should be taken in a responsible way, namely balancing the push/pull action (whole strategy and involvement) and enabling ethical perspectives on security enhancements.

CONCLUSION

COVID-19 caused the criticality of technology by increasing the use of e-learning. With the world moving towards being increasingly dependent on computers and automation, one of the main challenges in the current decade has been to build secure applications, systems, and networks. Therefore, in order to minimise potential failures, it is crucial to involve all stakeholders to have security awareness and security sensitivity as a prerequisite for trust and confidence in national and international successful e-learning implementation. Trust is confidence on someone's commitment to achieve a security goal.

Online learning is built on trust, information exchange, and discussion. However, due to the unexpected problem of the pandemic COVID-19, online learning providers had and still have to face a difficult balance, trying to provide sufficient security to protect online learning resources while not inhibiting the appropriate use of these resources. Our study focused on the suggestion of development and implementation of an improved e-learner model that supports monitoring of user behaviour related to information security. It is important to plan a risk mitigation in e-learning security.

KEYWORDS: e-learning, COVID-19, Security, Trust, Confidence.

REFERENCES

Almaiah, M. A., Al-Khasawneh, A., & Althunibat, A. (2020). Exploring the critical challenges and factors influencing the E-learning system usage during COVID-19 pandemic. *Education and information technologies*, 1(20). https://doi.org/10.1007/s10639-020-10219-y

Alomar, N.; Alsaleh, M.; Alarifi, A. (2017). Social authentication applications, attacks, defense strategies and future research directions: A systematic review. IEEE Commun. Surv. Tutor., https://doi.org/10.1109/COMST.2017.2651741

Arora, A. S., & Sharma, M. K. (2013). A proposed architecture of cloud computing based e-learning system. International Journal of Computer Science, & Network Security, 13(8), 31-34.

Asmaa, K., & Najib, E. (2016), E-learning Systems Risks and their Security. International Journal of Computer Science and Information Security (IJCSIS), 14(7).

Bandara, I.; Ioras, F. and Maher, K. (2014). Cyber Security Concerns in E-learning Education. In: Proceedings of ICERI2014 Conference, IATED, 0728-0734.

Bird, C., & Bhardwaj, H. (2020). From Crisis to Opportunity: Rethinking Education in the Wake of COVID-19, Child, & Youth Services, 41, 228-230.

Blinco, K., Mason, J., McLean, N., & Wilson, S. (2004). Trends and issues in e-learning infrastructure development. White Paper, ALT-I-LAB, California. Retrieved from http://www.jisc.ac.uk/uploaded_documents/Altilab04-infrastructureV2.pdf

Brammer, S., & Clark, T. (2020). COVID-19 and management education: Reflections on challenges, opportunities, and potential futures. British Journal of Management, 31, 453-456.

CVE (2021). Moodle: Vulnerability Statistics. Retrieved from https://www.cvedetails.com/product/3590/Moodle-Moodle.html?vendor_id=2105

EDRI. (2021). A victory for us all: European Court of Justice makes landmark ruling to invalidate the Privacy Shield. European Digital Rights Ass. Retrieved from https://edri.org/our-work/a-victory-for-us-all-european-court-of-justice-makes-landmark-ruling-to-invalidate-the-privacy-shield

E-learning Inside (2021). How COVID-19 Has Changed Education and How to Adapt. E-learning Inside, 08 January. Retrieved from https://news.elearninginside.com/how-covid-19-has-changed-education-and-how-to-adapt/ (25/03/2021)

Goyal, M., & Krishnamurthi, R. (2019). An Enhanced Integration of Voice-, Face-, and Signature-Based Authentication System for Learning Content Management System. In Kumar, A. (Ed.), Biometric Authentication in Online Learning Environments (pp. 70-96). IGI Global. http://doi:10.4018/978-1-5225-7724-9.ch004

Hamburg, I., & Grosch, K. R. (2017). Ethical Aspects in Cyber Security. Archives of Business Research, 5(10), 199-206.

Harini, N.; Padmanabhan, T.R. (2013). 2CAuth: A new two factor authentication scheme using QR-code. Int. J. Eng. Technol. 2013, 5, 1087–1094.

Hartley, R. (2014). Conceptualising and supporting the learning process by conceptual mapping. Smart Learn. Environ. 1(7). https://doi.org/10.1186/s40561-014-0007-2

Itani, W., Ghali, C., Kayssi, A., Chehab, A. (2014) Reputation as a Service: A System for Ranking Service Providers in Cloud Systems. In: Nepal S., Pathan M. (eds) Security, Privacy and Trust in Cloud Systems. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-38586-5_13

May, M., & George, S. (2011). Using students tracking data in E-learning: Are we always aware of security and privacy concerns 2011. IEEE 3rd International Conference on Communication Software and Networks: ICCSN 2011.

May, Madeth, Fesakis, Georgios, Dimitracopoulou, Angelique, & George, Sébastien. (2012). A Study on User's Perception in E-learning Security and Privacy Issues. Proceedings of the 12th IEEE International Conference on Advanced Learning Technologies, ICALT 2012. 88-89. https://doi.org/10.1109/ICALT.2012.145.

May, M., Iksal, S., Usener, C. A. (2017). The side effect of learning analytics: An empirical study on e-learning technologies and user privacy. In Costagliola, G., Uhomoibhi, J., Zvacek, S., McLaren, B. M. (Eds.), *Computers supported education* (Vol. 739, pp. 279–295).

Metared (2021). Kit de concienciación en ciberseguridad - Edición IES iberoamericanas. Retrieved from https://www.metared.org/global/kit-concienciacion-cyber.html

Miguel, J., Caballé, S., & Xhafa, F. (2017). Intelligent Data Analysis for e-learning: Enhancing Security and Trustworthiness in Online Learning Systems. San Diego: Academic Press.

Milošević, M., Milošević, D. (2016). Defining the e-learner’s security profile: Towards awareness improvement. Sādhanā 41, 317–326. https://doi.org/10.1007/s12046-016-0478-7

Mohsin, J.; Han, L.; Hammoudeh, M.; Hegarty, R. (2017). Two Factor vs. Multi-factor, an Authentication Battle in Mobile Cloud Computing Environments. In Proceedings of the International Conference on Future Networks and Distributed Systems, Cambridge, UK, 19–20 July 2017; ACM: New York, NY, USA, 2017; p. 39.

Moneo, J. M., Caballé, S., Xhafa, F., Prieto-Blázquez, J., & Barolli, L. (2016). A methodological approach for **trust**worthiness assessment and prediction in mobile online collaborative learning. Computer Standards, & Interfaces, 44, 122-136. https://doi.org/10.1016/j.csi.2015.04.008

NIST (2012). Computer Security Incident Handling Guide NIST. Retrieved from https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r2.pdf

Okafor, N. U., Oparah, C. C., & Okwudili, U.M. (2018). Security and Privacy in E-learning Education. Journal of Arts, Physical and Social Sciences, Federal Polytechnic Nekede, Owerri 1(1): 40-43.

Pearson, S. (2013). Privacy, security and trust in cloud computing. In S. Pearson, & G. Yee (Eds.), Privacy and Security for Cloud Computing (pp. 3-43). London: Computer Communications and Networks. Springer-Verlag.

Popel, M., & Shyshkina, M. (2019). The areas of educational studies of the cloud-based learning systems. Educational Dimension, 53(1), 60-79.

Rand Europe. (2011). The cloud: Understanding the security, privacy and trust challenges (technical report). RAND. Retrieved from http://www.rand.org.

Rawtani, M. R. (2012). Achieving knowledge management through cloud computing. In V. Prakash et al. (Eds.), *8th Convention PLANNER-2012* (pp. 387-394). Gangtok: Sikkim University.

Salter, J. (2019). Office 365 declared illegal in German schools due to privacy risks. Arstechnica. Retrieved from https://arstechnica.com/information-technology/2019/07/germany-threatens-to-break-up-with-microsoft-office-again/

Schinagl, S., Schoon, K., & Paans, R. (2015). "A Framework for Designing a Security Operations Centre (SOC)," 2015 48th Hawaii International Conference on System Sciences, 2015, pp. 2253-2262, doi: 10.1109/HICSS.2015.270.

Sharkey, A. J. C. (2016). Should we welcome robot teachers? Ethics and Information Technology, 18(4), 283–297. https://doi.org/10.1007/s10676-016-9387-z

Wani, F. H., & Khan, R. A. (2016). A Study of Security and Privacy Issues in E-learning Platforms. Two day national seminar on Electronic Devices, System and Information Security (SEEDS-2016) held by Department of Electronics, & IT, University of Kashmir from 18-19 March, 2016.

Weippl, E.R., 2005. Security in e-learning, New York, NY: Springer.

World Health Organization (WHO). (2020). Listings of WHO's response to COVID-19. Retrieved from https://www.who.int/news/item/29-06-2020-covidtimeline

Xie, X., Siau, K., Fui-Hoon Nah, F. (2020) COVID-19 pandemic - Online education in the new normal and the next normal. Journal of Information Technology Case an Application Research, 22, 175-187.

Zhang, D. (2004). Can e-learning replace classroom learning? *Communications of the ACM,* 47(5), 75-79.

Zhao, Y., Watterston, J. (2021). The changes we need: Education post COVID-19. Journal of Educational Change, 22, 3–12.

The ETHICOMP Book series fosters an international community of scholars and technologists, including computer professionals and business professionals from industry who share their research, ideas and trends in the emerging technological society with regard to ethics. Information technologies are transforming our lives, becoming a key resource that makes our day to day activities inconceivable without their use.
The degree of dependence on ICT is growing every day, making it necessary to reshape the ethical role of technology in order to balance society's 'techno-welfare' with the ethical use of technologies. Ethical paradigms should be adapted to societal needs, shifting from traditional non-technological ethical principles to ethical paradigms aligned with current challenges in the smart society.

UNIVERSIDAD DE LA RIOJA
Telefónica
Cátedra Telefónica. Aula Smartcities
UNIVERSIDAD COMPLUTENSE MADRID
UNIVERSITAT DE BARCELONA
UNIVERSITAT ROVIRA i VIRGILI