ARTIFICIAL INTELLIGENCE: HOW TO DISCUSS ABOUT IT IN ETHICS

Olli I. Heimo, Kai K. Kimppa

University of Turku (Finland), University of Turku (Finland)

olli.heimo@utu.fi; kai.kimppa@utu.fi

ABSTRACT

In this paper we look into how several different AI technologies are addressed in the ethics literature. We claim that in many cases the technologies are not defined well enough for the moral concerns to be as relevant as they could. We propose that for AI and ethics research to be taken seriously by those designing, using and creating policy, the ethical research to AI needs to be more specific on the technologies evaluated from an ethical perspective, and descriptive understanding of the technologies in question must be presented more clearly for the normative suggestions to be considered valid.

KEYWORDS: Artificial Intelligence, Ethics, Weak AI, Strong AI, Discourse.

1. INTRODUCTION

Artificial intelligence (AI) is the buzzword for the era and is penetrating our society in levels unimagined before – or so it seems to be (see e.g. Newman, 2018; Branche, 2019; Horaczek, 2019). In IT-ethics discourse there is plenty of discussion about the dangers of AI (see e.g. Gerdes & Øhstrøm 2015) and the discourse seems to vary from loss of privacy (see e.g. Belloni et al. 2014) to outright nuclear war (See e.g. Arnold & Scheutz 2018) in the spirit of the movie *Terminator 2*.

AI is presented sometimes as a bogeyman-technology, sometimes as a saviour of our age destined to save us from climate change, overpopulation, food shortage etc. Yet it seems that with AI discussion there is a lot of space for misunderstandings and misrepresentations starting from but not limited to what is AI. In this paper therefore the AI from the ethical perspective of what we should discuss about AI is presented.

This question will become more prevalent the more AI is being used in different circumstances. Actual applications behave very differently, even with same 'base' AI technology, depending on the application area, and even individual application. Thus, understanding and describing the application and the area for which the AI is being used as a solution becomes paramount to understand the specific ethical issues raised by the application; when there are ethical issues – not all applications of AI produce ethical concerns (e.g. using AI to separate different kinds of metal, wood and plastic from waste products), but rather only practical questions. Very high level attempts at ethical analyses will necessarily prove problematic; even military applications of AI can be ethically done, even if we would agree that AI automated weapons ought not to be created. Thus, the first step offered in this paper is to divide the area to different topic areas. In future papers, this division needs to be handled in more detail in each specific area, and those more specific areas need to be analysed in turn to find the areas with more and less ethical issues; although in the end, the question is always on an individual application level.

2. WHAT IS AI

2.1. General definition

There is of course various different ways to conceptualise the difference between different kinds of things labelled as AI. Whereas the technical ones have the tendency to focus on the technical structure of the tool at hand, from the ethical point of view the focus should be more on 1) what the system can do and 2) how it does it. Moreover, we should also focus on the issue on how the bad consequences could be avoided (Mill 1863) and how the people with malicious intentions could be controlled (Rawls 1971). There of course are different motivations and (hopeful) consequences when using AI, which are duly worthy of a different discourse and study in themselves), but in this paper the issue of *definition* for the use itself is discussed. Hence, in the full paper we will discuss the following four different groups of AI:

1. Scripts (gaming and otherwise)
2. Data mining and analysis
3. Weak AI & Strong AI (In its current form: neural networks, machine learning, mutating algorithms etc.)
4. General AI (Skynet, HAL, Ex Machina, etc.)

2.2. Scripts

First of all the *scripts*, mostly advertised as “AI” in computer games are just “simple” algorithms. As these are mostly the first version of AI we meet when talking about it, we must remember that they are merely scripts and cheating (i.e. not AI at all) to make the opponents in computer games more lifelike, to make the sensation that you are playing against actual intelligent opponents. This of course is not true because the easiest, cheapest, and thus most profitable way to give the illusion of a smart enemy is to give the script the power of knowing something they should not.

Hence the idea is to give the player the illusion, but the actual implementation is much simpler (and for smarter or more experienced players also quite transparent…). That is the art of making a good computer game opponent. Hence computer game AIs are just glorified mathematical models to entertain the customers.

2.3. Traditional data mining

The second one discussed as an AI quite often is *data mining* and the related data analysis, “just” gathering specific information from a huge pile of data. Data mining is “the science of extracting useful knowledge from such huge data repositories”(Chakrabarti et al., 2006). Yet this is usually and mostly done by scripting; Patterns and mathematical models are found and tiny bits of data from the patterns are combined to find similarities, extraordinarities and peculiarities then to be analysed by humans aided by a traditional algorithm. Data mining is a multidisciplinary field of study combining broadly statistics, linear algebra, database systems, and algorithms and data structures where the information stored can be made knowledge. (Chakrabarti et al., 2006, Hand, 2017.)

Traditionally there is nothing intelligent about these algorithms except the people making them. Therefore, compared to real artificial intelligence, they too are just glorified mathematical models and smart people working with them – a massive difference to the former though. It is of course possible, and in many cases advisable, to use machine learning and mutating algorithms in data mining (Wu, 2004), but as it is not required, in this categorisation, those deserve a place of their own.

2.4. Weak AI & Strong AI

Thirdly, we discuss machine learning, mutating algorithms, neural networks and other state of the art AI research, i. e. *weak AI*. This is *the point* we should currently focus on when discussing themes related to AI. These methods make the computer better by every step the computer makes; every decision the computer makes improves the computer, not the user.

To clarify, Artificial Intelligence refers to a system, in which is a mutating algorithm, a neural network, or similar structure (also known as weak AI) where the computer program "learns" from the data and feedback it is given. Weak AI is only capable on solving certain problems in chosen platforms and cannot achieve consciousness. It can although be rather excellent in identifying text, in speech-to-text applications, translation, identifying humans, human emotions, and actions from pictures and videos, and playing chess, go, checkers and other games. (Pietikäinen & Silvén, 2019pp. 23, 104-113)

Strong AI is an AI which is close on human intelligence and has at least some idea of self. The machine can use different background information while planning and making decisions. Fully autonomous actions in chaning environment, e.g. in traffic, already require partially a strong AI. Especially in conflict situations even though the lower level decisions, noticing other road users or chaning lines, are clearly in the territory of weak AI. To duplicate a natural and believable discussion between a human and a machine a strong AI is required due to the necessity to understand the context of the discussion. Strong AI is clearly the next big step in AI development. (Pietikäinen & Silvén, 2019, pp. 23, 113)

These technologies are usually opaque (i.e. black box –design), so even their owners or creators cannot *know* how or why the AI ended up with the particular end-result. (See e.g. Covington, Adams, and Sargin, 2016). As AI has been penetrating the society in many different levels for years, e.g. banking, insurance, and financial sectors (see e.g. Coeckelbergh, 2015).

2.5. General AI

The fourth issue, *General AI*, (sometimes *Artificial General Intelligence, AGI* (see e.g. Goertzel, 2007, p. V)), often discussed in the field of AI and described in multitude of Sci-Fi is the "living" AI, the thinking AI – possibly the feeling and fearing AI. The issue with a general AI is that we seem to be nowhere near in science. There are many "general AI" studies done in specific settings, e.g. gaming, where the development is focused in the AI learning to play different video games. These however are not general AIs as such, but moreover machine learning algorithms.

There is also a general AI category Super AI (also known as superintelligence), e.g. the "Skynet", the singularity "the moment at which intelligence embedded in silicon surpasses human intelligence" (Burkhardt, 2011, Pietikäinen & Silvén, 2019, pp. 23-24, Coeckelbergh , 2020, pp. 10-13) and starts to consider itself equal or better than humans. These AIs are luckily or sadly, depending on the narrative the utopia or the dystopia, are still mere fiction and in the technological scale in a future we cannot yet even comprehend.

3. PROBLEM

When discussing technology, the possibilities of technology and possible technologies we must be aware that the first of these does already exist. The second one of these is due to exist, and the third one may exist. While it is possible that technology will exist in say 5-10 years, we also must remember that the society will not be what it is now and other technologies will exist and the society has moved on. There are numerous issues within the field of AI currently at hand, e.g. biased AI (Heimo & Kimppa 2019), liability of autonomous vehicles (see e.g. Heimo, Kimppa & Hakkala 2019), weaponizing AI systems (see e.g. Gotterbarn, 2010), facial recognition (see e.g. , Heimo & Kimppa 2019; Doffman, 2019) just to mention few. Moreover there are plenty of near-future applications of these that must be handled before they become a critical issue. Yet it is important to discuss about all the levels of AI technologies – and to tie them to their timeline!

As we know we must interpret the writings of the past for they were written in their time (see e.g. MacIntyre, 2014), we must also interpret the future which will be different in ways we cannot fully understand. Therefore to predict the AI can do in 10-20 years' time is quite different when we cannot fathom what kind of society we will have in 10 years' time. We must yet keep in mind that what we give up now in the sense of privacy, personal information, liberties etc. can and will be taken away from us more efficiently with the future AI, especially if we follow the Chinese route, which is possible. But to talk of the society now with a futuristic AI seems intellectually dishonest. We do not have flying cars, hoverboards nor the cure for cancer, things predicted and assumed by everyone in any popular culture from the 80s or 90s (see e.g. Back to the Future) yet we have Twitter, Wikipedia and cat picture memes, not something we would actually have been predicting at the time. It is not that we would say that predicting future is irrelevant, moreover we wish to encourage people, scientists and philosophers to focus be explicit when predicting the future; to emphasize their predictions of the timeline they assume technology be in use. Hence when we are talking about AI there are many possibilities for the future but a General AI is a as much of a thing of a future we cannot yet predict, as datamining is a thing of the past. Predictions as predictions, and facts as facts, that is all we can do for honest science.

4. DISCUSSION

Therefore, when analysing digitalisation via AI and it's possibilities, it is clear that we should focus on weak AI and strong AI. These are the things of now and near future whereas scripts and data mining are not AI at all and general AI is still being sci-fi which we are not yet sure shall it happen, and if, when. Yet to create valid scientific discourse we should be focused on what we know instead of what we do not. To make predictions, alert other scientists (as it is a proper task for an IT-ethicist), and to guide the scientific discourse and development, we need that knowledge.

AIs already control a lot of our daily lives, e.g. in entertainment where Netflix, YouTube, and Social media sites which content is shown to us due our preferences, how we are classified by the system, and what the media corporation wishes to promote. This however seems to be still quite dumb and does not fulfil the promises marketed to the public. Since the algorithms generate frustration in the users due poor suggestions as majority of the content one wishes to see must be acquired by searching. Yet the AIs are learning and might turn out to be the privacy endangerment predicted. (Heimo & Kimppa, 2019)

One of the key issues when talking AI is the black box –mentality of the given systems. Whereas we can understand where our solutions come from and tweak them to be ethical (e.g. not discriminatory against women, as was the case in Amazon's HR (Hamilton, 2018)). The black box feature is one of the key issues when discussing about the AI in ethics and a key reason why the definitions around AI should be clearly expressed.

Also the question of when is important. As focusing on the discussing about AI, the distinction between now, near-future and far-future should be made clear. If the discussion around time-frame obscures, the discussion itself can become obscured due to the predicted development of technology and the various other possibilities in the future. Therefore, if the time-frame of the discussion is not clear, the discussion is no longer valid as we are not discussing about the same thing anymore.

Hence the authors propose a two-stage model on evaluating AI:

- Are we talking about AI or something else? Describe the AI clerly.
- Are we talking now/near-future or far-future/sci-fi? Tell the audience roughly the time-frame, e.g. 5-10 years or 30-50 years.

A fine example on the discourse without timelines is in Coeckelbergh's (2010) esteemed article "Robot rights? Towards a social-relational justification of moral consideration" where Coeckelbergh, rises interesting arguments about robot rights and finds equally interesting questions and justifications. Yet the article lacks the depth in the description of AI development timelines on predicting the need for the change mentioning only "near-future" and "long stage", which after 10 years seem to be still that. The main goal of this article of course is not to alarm us to the imminent requirement for robot rights nor demand any action for or against the current development but moreover to participate to an academic discourse presented in the paper.

Yet the argument of this paper is that we should improve the precision when discussing future technologies – especially *with near-future applications* and at least *when rising alarm or demanding action*. The prediction of this paper is that the future of predicting future is danger if the current predictions of future are done without clearly describing the foreseeable future.

5. CONCLUSIONS

What we want to emphasize with this paper is that many authors on the ethics of AI leave the kind of AI they are discussing so unclear as to not make it clear whether they even understand the topic area at all. They have vague notions of AI, which they do not specify to the extent that the ethical questions are first of all not relevant to any specific technology currently used, nor clearly future studies on the problematic paths that we may take. This causes AI ethics not to be taken seriously by those who ought to take it seriously, namely designers of AI, companies using AI, and governments and intergovernmental organizations attempting to regulate AI development.

If we in the field of ICT and ethics are not believable, our suggestions will be ignored, and AI development may be either misdirected or left all together undirected, and thus create applications which are problematic for users, companies and the society alike. Especially considering the surprising amount of AI ethicists that have recently emerged from anonymity on the field, traditional ICT and ethics researchers who have done years, even decades of study in the field of AI and ethics need to be extremely careful to see to the validity of their claims, whilst at the same time they need to be very visible in the current discussions relating to AI and ethics in all relevant levels from concept creation to actual applications to government and intergovernmental policy creation.

REFERENCES

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.

Belloni, A. et al. (2014) Towards A Framework To Deal With Ethical Conflicts In Autonomous Agents And Multi-Agent Systems, CEPE 2014.

Branche, P. (2019), Artificial Intelligence Beyond The Buzzword From Two Fintech CEOs, Forbes, Aug 21 2019, https://www.forbes.com/sites/philippebranch/2019/08/21/artificial-intelligence-beyond-the-buzzword-from-two-fintech-ceos/#43f741c7113d

Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., ... & Wang, W. (2006). Data mining curriculum: A proposal (Version 1.0). Intensive Working Group of ACM SIGKDD Curriculum Committee, 140.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. Ethics and information technology, 12(3), 209-221.

Coeckelbergh, M. (2015) The tragedy of the master: automation, vulnerability, and distance, Ethics and Information Technology, 17:219-229.

Coeckelbergh, M. (2020) AI Ethics, MIT Press, 2020.

Covington, P., Adams, J., and Sargin, E. (2016) Deep neural networks for youtube recommendations. Proceedings of the 10th ACM conference on recommender systems. ACM, 2016.

Doffman, Z. (2019) China’s ‘Abusive’ Facial Recognition Machine Targeted By New U.S. Sanctions, Forbes, Oct 8, 2019. https://www.forbes.com/sites/zakdoffman/2019/10/08/trump-lands-crushing-new-blow-on-chinas-facial-recognition-unicorns/#52641d79283a

Gerdes, A. & Øhstrøm, P. (2015) Issues in robot ethics seen through the lens of a moral Turing test, JICES 13/2:98-109.

Goertzel, B. (2007). Artificial general intelligence (Vol. 2). C. Pennachin (Ed.). New York: Springer.

Gotterbarn, D. (2010) Autonomous weapon's ethical decsions:" I am sorry Dave; I am afraid I can't do that.". In proceedings of ETHICOMP 2010 The “backwards, forwards and sideways” changes of ICT Universitat Rovira i Virgili, Tarragona, Spain 14 to 16 April 2010.

Hamilton, I.A. (2018) Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women, Business insider, October 10th 2018, available at https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10?r=US&IR=T

Hand, D. J. (2007). Principles of data mining. Drug safety, 30(7), 621-622.

Heimo, O. I. & Kimppa, K. K. (2019) No Worries–the AI Is Dumb (for Now), Proceedings of the Third Seminar on Technology Ethics 2019 Turku, Finland, October 23-24, 2019, pp. 1-8.

Heimo, O. I., Kimppa, K. K. & Hakkala, A (2019) Automated automobiles in Society, IEEE Smart World Congress, Leicester, UK, 2019.

Horaczek, S. (2019), A handy guide to the tech buzzwords from CES 2019, Popular Science Jan 9 2019, https://www.popsci.com/ces-buzzwords/

Mill, John S. (1863) Utilitarianism, https://www.utilitarianism.com/mill1.htm, accessed 21.10.2019.

Newman, D. (2018) Top 10 Digital Transformation Trends For 2019, Forbes, Sep 11, 2018, https://www.forbes.com/sites/danielnewman/2018/09/11/top-10-digital-transformation-trends-for-2019/#279e1bca3c30

Pietikäinen, M. & Silvén, O. (2019) Tekoälyn haasteet – koneoppimisesta ja konenäöstä tunnetekoälyyn, Center for Machine Vision and Signal Analysis (CMVS), November 2019, ISBN 978-952-62-2482-4

Rawls, J. (1971) A Theory of Justice, Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Robbins S. (2018) The Dark Ages of AI, Ethicomp 2018.

Wu, X. (2004) Data mining: artificial intelligence in data analysis. In Proceedings. IEEE/WIC/ACM International Conference on Intelligent Agent Technology, 2004.(IAT 2004). (p. 7). IEEE.