EMBEDDING VALUES IN AI BY DESIGN: AN INTEGRATED FRAMEWORK

Xenia Ziouvelou, Vangelis Karkaletsis, Konstantina Giouvanopoulou

AI Politeia Lab, National Centre of Scientific Research "Demokritos" (Greece)

xeniaziouvelou@iit.demokritos.gr; vangelis@iit.demokritos.gr; kgiouvano@iit.demokritos.gr

EXTENDED ABSTRACT

Artificial Intelligence (AI) is evolving rapidly, becoming a key driver for the digital transformation of our economies and societies. Impacting this way the future of humanity, by transforming the lives of individuals and influencing human societies, reshaping patterns of living, working, learning and interacting. However, while AI can create great opportunities by driving economic and social progress, it also presents complex challenges and potential risks. Risks are related to gender-based or other kinds of discrimination and bias (intentional and unintentional), opaque decision-making, intrusion, social harms for individuals and society, loss of liberty, control and autonomy, in addition to the concentration of power in the hands of a few private actors, among others (UNESCO, 2020; EIGE, 2021). Challenges on the other hand, stem from the great uncertainties that are linked with the alignment of AI systems with human values (AI value alignment) from their design to their use (Han et al., 2022); which is of major concern given that the way we model and design AI may affect the values we are able to embed (Gabriel, 2020; Van de Poel, 2020). However, there are other dimensions. The evolution of AI brings about the need to explore deeper the interplay between values and technology design, development, implementation, and use and the role of individuals in realising value sensitive technology; as well as the need explore new values, which are appropriate to protect the rights of the individual in the light of such an evolution (Ziouvelou et al., 2020).

Beyond these anticipated risks, there are increasing concerns over unintended and unanticipated risks with negative, undesirable impacts that may accompany AI technology and its applications. Triggered by these risks, a growing body of ethical AI guidelines and principles, has emerged over the last few years (Hagendorff, 2020, 2022, EU HLEG, 2019; Whittaker et al., 2018; Campolo et al., 2017; Floridi et al., 2018; IEE, 2019; Jobin et al., 2019; among others) aiming to harness the unintended disruptive potential and complex challenges posed by AI. Numerous guidelines have been launched by governments, scientific or industrial communities as well as civil society representatives, aiming to serve as a basis for ethical decision-making in AI design, development, deployment and governance. However, public debate is already saturated by these ethical guidelines. From a macroscopic perspective, this abundance of ethical principles threatens on the one side to overwhelm and confuse and on the other to delay the development of laws, rules and standards that will ensure that AI is socially beneficial (Floridi and Cowls, 2019) or even avoid regulation altogether (Wagner, 2018) in some geographical regions. From a microscopic perspective, the vast majority of these guidelines appear to adopt the 'deontological ethical approach' (Mittelstadt et al., 2019; Hagendorff, 2020), that emphasises duties or rules at an institutional level. At an individual level though, there appears to be a gap, for example in relation to the values, moral and character dispositions of the individuals who create these technologies (at an individual and

company level). Business, government and civil society leaders need to understand the importance of values and ethics in technological development in order to seize the opportunities and address the threats that accompany emerging technologies (seen as sociotechnical systems rather than isolated artifacts (Van de Poel, 2020)), and this implies adopting a conscious perspective on technological development that prioritises society's values (Philbeck et al., 2018). As such, virtue ethics could expand traditional deontological AI ethics and broaden the scope of action (Hagendorff, 2020, Van de Poel, 2020).

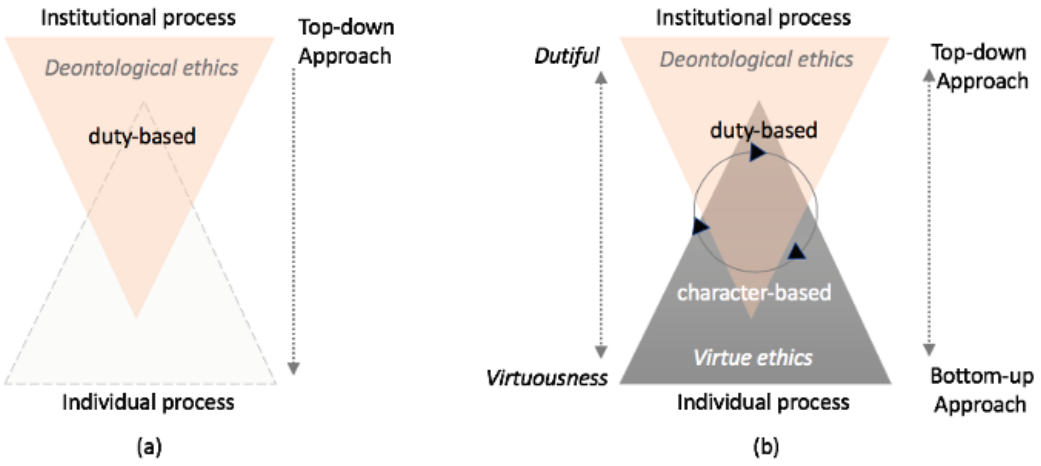
In this paper, we map the existing landscape of systematic scoping studies in the area of ethical guidelines for AI and we examine whether convergence is emerging in relation to the principles, Furthermore, we examine the theoretical parameters underpinning these ethical guidelines, identify gaps and provide a complementary perspective that aims to provide an integrated, multiperspective approach to the discussion about what constitutes 'ethical AI '. We support the view that understanding human values is a crucial step in the development of responsible and trustworthy AI (Han et al., 2022). Our perspective is anchored in the analysis of Hagendorff (2020) and Van de Poel (2020) and the need for a holistic framework for AI ethics that will present a model that augments the traditional, prevalent deontological approach of AI ethics (Mittelstadt et al., 2019; Hagendorff, 2020) (Figure 1a) with an approach oriented towards virtue ethics pertaining values, moral and character dispositions (Van de Poel, 2020). As it is very difficult to predict exactly what kind of consequences future innovations will bring to society, Shannon Vallor (2016) argues for virtue ethics as an appropriate framework for the development of emerging technologies, which, by enabling new forms of behaviour, are expected to influence human values in the future (Steen et al., 2021).

The difference between deontology and virtue ethics is that while the former is based on normative rules with universal validity, the latter examines what constitutes a good person or character. Ethics focuses on the act, while virtue focuses on the actor, on the development of positive characteristics of the actor (Hagendorff, 2020) and is essential if we consider that the values that every individual engineer embraces should be the starting point for responsible and ethical behaviour (Hersh, 2012). Values contribute to evaluation in terms of goodness and badness, while concepts such as duties and rules are used to determine the rightness (or wrongness) of actions (Van de Poel, 2020). The virtuous actor embraces values of goodness, therefore the ethics of virtue is directly related to moral values. As stated by Annas (2011) virtue is a disposition of character, which is not impermanent, to act reliably and virtue requires commitment to values, it involves the orientation of the person to something that the person considers valuable. In an effort to make the implementation of existing AI ethics initiatives successful and effective, the insights of moral psychology should be included, since until now, when talking about AI ethics, the psychological processes that limit the goals and effectiveness of ethics programs are not taken into account (Hagendorff, 2020).

Our motivation is in developing a model that will adopt such an integrated approach to AI ethics that will augment the existing duty-driven approach including principles and rules (i.e., ethical AI code) (Figure 1a - deontological approach) with a virtue-driven approach including values and moral personality traits (i.e., moral/value AI code) (Figure 1b- integrated approach). This model will thus, broaden the scope of action by infusing virtues and ethos in AI ethics. Ethos means ''virtue" (the translation of the Ancient Greek word ἀρετή - ''arete'') in the Aristotelian sense and denotes the internal values that characterise an individual. Aristotle argued that man is by nature zoon politikon, destined to live in an organised political society

and that virtues contribute to living in a polis and promoting the welfare of the people (Steen et al., 2021). The word virtue denotes moral excellence, and it indicates the fundamental qualities that allow people to excel and thus contribute to social well-being. Virtue ethics is a theory that although it cannot guarantee beneficial societal innovation, it can nevertheless be of value in a holistic framework that complements the existing deontological ethics, as illustrated in the figure 1.

Figure 1: (a) Deontological AI ethics approach & (b) Integrated AI ethics framework (Deontological & Virtue ethics).



Considering that artificial systems are not able to understand the notion of human values (Neuhäuser, 2015), lack emotional abilities (Sharkey, 2017) and are human made (Hakli & Mäkelä, 2019), all concern should be about humans involved in designing, developing and deploying AI systems. Given that humans can be considered full moral agents (Dignum, 2018; Hakli & Mäkelä, 2019), virtue ethics suggests that we do not treat AI systems as autonomous, equal to humans, but rather as assistive companions (Maes, 1995; Savulescu & Maslen, 2015; Voinea et al, 2020) as intelligent tools (Balkin, 2017) to serve human needs in a responsible way. Furthermore, just as individuals can demonstrate character, an organization can also embody character, as a collection of individuals (Moore, 2005). Existing research indicates that organizations that demonstrate virtue by exhibiting character, tend to experience positive benefits both internally as well as in the marketplace (Cameron, Bright, and Caza, 2004; Sosik, Gentry, and Chun, 2012; Neubert and Montañez, 2020).

This paper, aims to address the need for a holistic framework for AI ethics by design; a framework that will augment the current prevalent approach with a virtue-driven approach aiming at values, moral and character dispositions of individuals (human embedding values in AI systems (at an individual level and organizational level). To this end it provides an integrated approach to AI ethics aiming to broaden the scope of action by embedding virtues, ethos and values in AI by design. As such there is a need to explore the practical implementation of such a holistic approach and examine how the proposed framework can be implemented so as to foster responsible and trustworthy AI Systems based on an integrated ethics approach, that augments the current deontological approach by using virtue ethics. This will in turn provide

some useful insights into the interplay between virtue ethics and deontological ethics in the context of AI systems and values.

KEYWORDS: Artificial Intelligence, Values, Sociotechnical AI systems, Value embedding, Value-driven AI, AI Ethics Framework.

AKNOWLODGEMENTS: This research is funded by the project AI4EUROPE, Grant Agreement No 101070000 (EU funded project), under the HORIZON.2.4.5 - AI and Robotics.

REFERENCES

Annas, J. (2011). *Intelligent Virtue*. Oxford University.

Balkin, J. M. (2017). The three laws of robotics in the age of big data. *Ohio State Law Journal,* 78(5), 1217–1241.

Cameron, K. S., Bright, D., & Caza, A. (2004). Exploring the relationships between organizational virtuousness and performance. American Behavioral Scientist, 47(6), 766-790.

Campolo, A., Sanfilippo, M. R., Whittaker, M., & Crawford, K. (2017). *AI Now 2017 Report*. AI Now Institute at New York University.

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology, 20*(1), 1–3.

EIGE (2021). Artificial intelligence, platform work and gender equality. *European Institute for Gender Equality.* Luxembourg: Publications Office of the European Union.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., et al. (2018). AI4People - an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines* 28 (4), 689–707.

Floridi, L. and Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*.

Gabriel, I. (2020). Artificial Intelligence, Values and Alignment. *Minds and Machines, 30,* 411-437.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines,* 1-22.

Hagendorff, T. (2022). A Virtue-Based Framework to Support Putting AI Ethics into Practice. *Philos. Technol.* 35, 55.

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist, 102*(2), 259–275.

Han, S., Kelly, E., Nikou, S. & Svee, E.Q. (2022). Aligning artificial intelligence with human values: reflections from a phenomenological perspective. *AI & Soc* 37, 1383–1395.

Hersh, M.A. (2012). Science, Technology and Values: Promoting Ethics and Social Responsibility, *IFAC Proceedings Volumes, 45* (10), 79-84.

HLEG, (2019). A definition of AI: Main capabilities and disciplines, High-Level Expert Group on Artificial Intelligence of the European Commission. *Downloded,* 1, 2019-12.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature*

Moore, G. (2005). Corporate character: Modern virtue ethics and the virtuous corporation. Business Ethics Quarterly, 15(4), 659-685

Maes, P. (1995). Artificial life meets entertainment: Lifelike autonomous agents. *Communications of the ACM, 38*(11), 108–114.

Mittelstadt, B., Russell, C., and Wachter, S. (2019). *Explaining explanations in AI*. In Proceedings of the conference on fairness, accountability, and transparency—FAT* '19, 1–10.

Neubert, M. J., and Montañez, G. D. (2020). Virtue as a framework for the design and use of artificial intelligence. Business Horizons, 63(2), 195-204.

Neuhäuser, C. (2015). Some Skeptical Remarks Regarding Robot Responsibility and a Way Forward. In C. Misselhorn (Ed.), *Collective Action and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation* (pp. 131–146). Springer.

Philbeck, T., Davis, N. and Engtoft Larsen, A. M. (2018). White Paper. Values, Ethics and Innovation Rethinking Technological Development in the Fourth Industrial Revolution. *World Economic Forum.*

Savulescu, J., & Maslen, H. (2015). Moral Enhancement and Artificial Intelligence: Moral AI? In J. Romportl, E. Zackova, & J. Kelemen (Eds.), *Beyond Artificial Intelligence. The Disappearing Human-Machine Divide* (pp. 79–95). Springer.

Sosik, J. J., Gentry, W. A., & Chun, J. U. (2012). The value of virtue in the upper echelons: A multisource examination of executive character strengths and performance. The Leadership Quarterly, 23(3), 367-382.

Sharkey, A. (2017). Can robots be responsible moral agents? And why should we care? *Connection Science, 29*(3), 210–216.

Steen, M., Sand, M. & Poel, I. (2021). Virtue Ethics for Responsible Innovation. *Business & Professional Ethics Journal.*

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems.*

UNESCO (2020). Artificial Intelligence and Gender Equality, Key findings of UNESCO's Global Dialogue.

Vallor, S. (2016). Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. New York: Oxford University Press.

Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, *30*(3), 385-409.

Voinea, C., Vică, C., Mihailov, E., & Savulescu, J. (2020). The Internet as Cognitive Enhancement. *Science and Engineering Ethics,* 26(4), 2345–2362.

Wagner B. (2018). *Ethics as an escape from regulation: From "ethics-washing" to ethics-shopping?* In Bayamlioglu, E., Baraliuc, I., Janssens, L. A. W. & Hildebrandt, M. (Eds.). Being Profiled: Cogitas Ergo Sum: 10 Years of Profiling the European Citizen (pp. 84-89). Amsterdam: Amsterdam University Press.

Wallach, W. (2004). Artificial Morality: Bounded Rationality, Bounded Morality and Emotions. In I. Smit, G. Lasker and W. Wallach, editors, *Proceedings of the Intersymp 2004 Workshop on Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence,* pp. 1 – 6, Baden-Baden, Germany, IIAS, Windsor, Ontario.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., Schwartz, O. (2018). AI Now report 2018.

Ziouvelou X., V. Karkaletsis, G. Giannakopoulos, A. Nousias, S. Konstantopoulos (2020). Democratising AI: A National Strategy for Greece. NCSR Demokritos http://democratisingai.gr/