MAPPING AI ETHICS: FROM PRINCIPLE INTO PRACTICE

OBJECTIVE

The objective of this research proposal is to advance in closing the gap between practice and research in the area of AI/MACHINE ETHICS through a mapping of state of the art for Ethics in AI at different levels: macro, mezzo and micro:

- "Macro" (regional, national and significant corporation guidelines), which affect the "purpose", type of applications that might be allowed to use (or ban) in public spaces, for example, where privacy and security issues are paramount.

- "Meso" (applications/use cases,) Scenarios where there is a clear human-machine interaction: business decision making, recruitment, even autonomous cars, how is that hand-over of moral control between human-machine process managed? Is the person readily prepared for "assuming" control or taking the decision? Is the info for the decision "biased"? Can we expect a better decision? Which ethical aspects/framework have been considered?

- "Micro" (algorithmic implementations). Which are the current implementations of ethical framework/theories in machine learning or alternative AI models? Where are the current limits?

In terms of methodology, a literature review on the latest developments of AI/machine ethics at the three levels has been carried over.

INTRODUCTION

The research field of Ethics in Artificial Intelligence (AI) has attracted much interest recently (Jobin et al., 2019). Despite some debate as to whether moral agency concepts

might be applied to Artificial Intelligence (Camacho et al., 2019; Charisi et al., 2017), there is a growing consensus that digital technologies are legitimate objects of ethical concern (Greene et al., 2019), moving away from the technological neutrality view of the last decade. Intention, purpose and human values are embedded in the design of Artificial Intelligent systems, and therefore they are subject to ethical reasoning. In fact, there is an urgent and real need for a functional system of ethical reasoning as AI systems are ready to be deployed at a massive scale (Charisi et al., 2017).

There is also a need for Ethical Governance of AI Systems (Winfield and Jirotka, 2018). The purpose of these governance systems should be to generate adequate principles and standards, and to foster ethical behaviour in both individual designers and the organisations in which they work. There remain however several challenges for that development: defining and formalising the ethical issue (philosophic), implementing some degree of moral reasoning in autonomous systems (engineering), and connecting the outcome of such systems with real actions affecting business, people and society (decision making) (Boddington 2017; Winfield and Jirotka, 2018).

AI ethics is generally concerned with how the AI industry should behave in order to minimise the ethical harms that can arise from AI or, less frequently mentioned, how to maximise its potential benefits. This concern has already led to the development of ethical principles and guidelines. (Winfield et al., 2019). Machine Ethics may be considered a subdomain within AI ethics, focused on how to implement ethics in AI systems in a practical way (Charisi et al., 2017). This field spans philosophy, business and engineering areas, and therefore AI engineers need to engage more with the ethics and decision-making communities (Yu et al., 2018); together, they can leverage their expertise in the development of more ethical AI technologies.

FILLING THE GAP FROM THEORY INTO PRACTICE

From a philosophical point of view, AI ethics raises several questions: is the concept of moral agency strictly correct when applied to IA? When, and how can we assume the conditions of autonomy, intelligence and free will? (Boddington 2017). When is the moral control lost? (Camacho et al., 2019). Which is the motivation of an artificial agent to behave ethically? (Charisi et al., 2017).

Moor (2006) established a distinction between implicit ethical agents, that is machines designed to avoid unethical outcomes, and explicit ethical agents, that is machines which either directly encode or learn ethics and determine actions based on those ethics.

Several ethical theories have been applied to AI ethics: normative ethics (consequentialism, deontology and virtue ethics) (Yu et al., 2019; Carter et al., 2017), Rawls' veil of ignorance (Boyles, 2018), or Habermas' discourse ethics (Mingers and Walsham, 2010). However, there is no overall agreement on which ethical theory to apply or how to implement those.

Thanks to the efforts of several initiatives, the IA community is finding some degree of global convergence around five ethical principles: transparency, justice and fairness, non-maleficence, responsibility and privacy (Jobin et al., 2019). However, there is also a global claim for adequate implementation strategies and how to translate principles into practice (Morley et al., 2019; Vakkuri et al., 2019; Yu et al., 2018; Winfield and Jirotka, 2018)

There are generally two approaches to implementing ethical behaviour in machines (Winfield and Jirotka, 2018; Wallach and Allen, 2008). A constraint-based approach (also known as top-down or rule-based), explicitly constraining the actions of an AI system under certain moral norms; and a training approach (also known as bottom-up or example-based), allowing the AI system to be trained to recognise and correctly

respond to morally challenging situations. There might also be considered a mixed approach in which an AI system starts with a set of rules or values and modifies them into a system for discerning right from wrong. (Charisi et al., 2017).

While organisation level policies and guidelines can direct development work, microlevel decisions are nonetheless left to individual developers, and therefore a developer-centric approach to ethics in AI is essential (Vakkuri et al., 2019). Developers working with AI need to be able to implement ethics into the systems they develop. Recent work is being done on developing generalisable individual ethical decision frameworks combining rule-based and example-based approaches to resolving ethical dilemmas (Yu et al., 2018). Some models have been developed to consider data-driven examples (Balakrishnan et al., 2019), to reflect subjective preferences and ethical boundaries (Rossi and Mattei, 2019; Loreggia et al., 2018), to represent ethical dilemmas (Anderson and Anderson, 2014), Ethics Shaping, as a proposal to make reinforcement learners not only achieve the expected performance and the goals but also comply with ethical rules, using reward shaping and stochastic policy from human data (Wu and Lin, 2018), or even a software "exoskeleton" that enhances and protects users by mediating their interactions with the digital world according to personalised data (Autili et al., 2019).

However, there are many challenges still to be resolved, such as how to compare preferences and ethical boundaries, and how to combine them, how to approach ill-defined problems and messes (Hester & Adams, 2017), or how to design systems for multiple AI agents working together -or in conjunction with humans (Rossi and Mattei, 2019),

As mentioned above, a whole body of ethical guidelines has been developed in recent years. However, the answer to the question as to whether those ethical guidelines have an actual impact on human decision-making in the field of AI and machine learning is often unfavourable (Hagendorff, 2019).

There is little research on how people translate predictions into actionable decisions related to AI (Morley et al., 2019). However, AI systems are increasingly supporting human decision making, or they make decisions autonomously (Rossi and Mattei, 2019).

Most of the latest research and industry efforts have been towards analysing and avoiding data bias and developing fairness algorithms (Caliskan et al., 2017; Friedler et al., 2018, AI Now, 2018), but fairness does not equal ethical, at least not alone, and there are additional problems related to the correlation vs causation relationship (McQuillan, 2018).

Regarding decision-making, there are several scenarios: a) individual ethical decision frameworks; b) collective ethical decision frameworks; and c) ethics in human-AI interactions (Yu et al., 2018). There is a need for research in the area of business processes, data governance and decision making (Abrams et al., 2019).

References

Abrams, M., Abrams, J., Cullen, P., & Goldstein, L. (2019). Artificial Intelligence, Ethics, and Enhanced Data Stewardship. IEEE Security & Privacy, 17(2), 17-30.

AI NOW Report 2018, Artificial Intelligence Institute. New York.

Anderson, M., & Anderson, S. L. (2014, June). GenEth: A general ethical dilemma analyser. In Twenty-Eighth AAAI Conference on Artificial Intelligence.

Autili, M., Di Ruscio, D., Inverardi, P., Pelliccione, P., & Tivoli, M. (2019). A Software Exoskeleton to Protect and Support Citizen's Ethics and Privacy in the Digital World. IEEE Access, 7, 62011-62021.

Boddington, P. (2017). Towards a Code of Ethics for Artificial Intelligence, Springer.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." Science 356.6334 (2017): 183-186. http://opus.bath.ac.uk/55288/

Camacho, J, Gonzalez Fabre, R. and Tejedor, P. (2019). *Moral control and ownership in AI systems*. Manuscript submitted for publication.

Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., ... & Yampolskiy, R. (2017). Towards moral autonomous systems. arXiv preprint arXiv:1703.04741.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2018). "A comparative study of fairness-enhancing interventions in machine learning". arXiv preprint arXiv:1802.04422.

Greene, D., Hoffmann, A. L., & Stark, L. (2019, January). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. In Proceedings of the 52nd Hawaii International Conference on System Sciences.

Hagendorff, T. (2019). The Ethics of AI Ethics--An Evaluation of Guidelines. arXiv preprint arXiv:1903.03425.

Hester, P. T. and K. M. Adams (2017). Systemic Decision-Making Fundamentals for Addressing Problems and Messes. Springer.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

Loreggia, A., Mattei, N., Rossi, F., & Venable, K. B. (2018, March). Preferences and ethical principles in decision making. In 2018 AAAI Spring Symposium Series.

McQuillan, D. (2018). "People's councils for ethical machine learning". Social Media+ Society, 4(2), 2056305118768303.

Mingers, J., & Walsham, G. (2010). Toward ethical information systems: the contribution of discourse ethics. Mis Quarterly, 34(4), 833-854.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. IEEE intelligent systems, 21(4), 18-21.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices. arXiv preprint arXiv:1905.06876.

Rossi, F., & Mattei, N. (2019, July). Building ethically bounded AI. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 9785-9789).

Vakkuri, V., Kemell, K. K., Kultanen, J., Siponen, M., & Abrahamsson, P. (2019). Ethically Aligned Design of Autonomous Systems: Industry viewpoint and an empirical study. arXiv preprint arXiv:1906.07946.

Winfield, A. F., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180085.

Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). Machine ethics: the design and governance of ethical AI and autonomous systems. Proceedings of the IEEE, 107(3), 509-517.

Wu, Y. H., & Lin, S. D. (2018, April). A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents. In Thirty-Second AAAI Conference on Artificial Intelligence.

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into artificial intelligence. arXiv preprint arXiv:1812.02953.