Autonomous yet moral machines

Marija SLAVKOVIK,
University of Bergen, Norway,
marija.slavkovik@uib.no

Abstract. Autonomous machines, both software and embodied artificial intelligent (AI) agents, will continue to relieve us of the burden that are monotone, repetitive, or dangerous tasks. The more autonomous machines inhabit our everyday life, the more we need to concern ourselves with the extent of disruption that this technology ushers. This is a brief overview of the different fields tackling ethical concerns with respect to AI research and applications

1. Background

Recent advances in applications of artificial intelligence (AI), specifically, in machine learning and deep neural networks, can create the illusion that AI is a recent technological advance. AI has been an established research field since 1956 [22], but ethical concerns with respect to AI have only gotten a serious scientific attention since the begining of this century [19,28,4]. What has changed is not the existence, or reality, of AI applications, but who the users and environment of these AI applications are today [9].

Last century AI applications are either powerful enough to inflict physical harm to people, or "toys" that have very limited range of abilities, such as the autonomous vacuum cleaners [25]. Further more, the first category has been limited for use by specially trained professionals and restricted in movement to a so called *working envelope* - constrained space that only trained specialists are allowed to enter. Examples of this category of machines are industrial robots [20], automated subway systems, which have been in operation for the past forty years[1], complex scheduling systems using constraint satisfaction programming [24], etc. A less obvious example are autonomous intelligent software solutions such as financial management decision aid systems [30]. Such decision aid systems are purpose built to be used by professionals.

In contrast today, we not only have AI applications that are powerful enough to inflict pain and not restricted to a controlled environment, but also software and hardware that is capable of a certain degree of autonomy and is available off-the-shelf to be used in virtually infinite contexts of operation. Examples include the ubiquitous self-driving cars [17], but also smart software and devices [18], such as smart speakers [16]. Decision aid systems have also evolved to include prediction of future behavior based on data that may not have been specifically collected by domain experts for this purpose [14].

[1] `http://www.railjournal.com/index.php/metros/uitp-forecasts-2200km-of-automated-metros-by-2025.html`

2. What do we talk about when we talk about AI ethics?

Unlike any other technology, AI raises special societal ethical concerns by virtues of being capable of, however limited, autonomy and unconstrained interaction with people and property.

Responsible AI, but also accountable and transparent AI ,[11,10,23,26], concerns ethical issues of using AI methods or conducting AI research. As any other powerful tool that extends human abilities, AI can be used to benefit people, but also due to malice or negligence it can cause harm. Responsible AI is not only concerned with ensuring the ethics of AI research and applications as a whole. It is also concerned that the right mount of attention is used to ensure that the AI application does not build upon or propagate existing unethical practices in our society.

Explainable AI [1,21] is concerned with ensuring that the decisions and actions of an autonomous systems have explanations comprehensible for the people who use those decisions, coordinate with those actions, or are affected by them. The explainability of a system is particularly a concern in the use of those machine learning methods, such as deep learning, for which there is no built in clear or transparent way to explain why a particular prediction is made given particular data. To a certain extent, explainable AI is also about ensuring that a human is always kept "in the loop" as a safe-guard that human intentions and AI actions are aligned [29].

Machine ethics or artificial morality [19,28,4,9] is explicitly concerned with the implementation of moral behavior in autonomous systems. In ethically sensitive situations, where the system makes decisions and actions that require a sensitivity to right and wrong, we need to make sure that the machine has such sensitivity or is somehow constrained from choosing an unethical option.

3. What is a moral machine?

Human activities, regardless of how monotonous and repetitive they are, consist not only of a task, but also of mindfulness who we are as an agent executing that task and how our actions affect the environment in which they are executed. We do not seize to be a moral agent regardless of what we are doing. What does it mean for a person to be a moral agent and her actions to be good or bad, is the domain of moral philosophy [15].

When a machine is programmed to replace a human activity, explicit concerns have to be taken not to engineer away the ethical impact that activity has on its environment. In the past, when AI applications have been implemented in controlled operated environments or designed for trained users, these ethical concerns could be engineered by predicting and constraining any context in which they might arise. In today's applications, this approach is no longer viable.

While the question of can a machine be a moral agent is still open among moral philosophers [2,13], it is slowly becoming clear that in machines we speak of degrees of moral sensitivity. Four categories of ethically sensitive agents have been forward by [19]: ethical impact agents, implicit moral agents, explicit moral agents and full moral agents. Wallach and Allen [27, Chapter 2] distinguish between operational morality, functional morality, and full moral agency. The last category in both works refers to human level of morality, what ever that may mean.

Agents with ethical impact are autonomous systems that do not make ethical choices. Predicting whether a convict is likely to re-offend if released on parole is not a moral decision. However, these systems do change the moral fabric of the environment in which they exist. A system that predicts re-offense, or a system that recommends whether you should be given a loan, can do wrong by perpetuating a bias against a certain group of citizens [6].

Implicit moral agents and agents with operational morality are autonomous systems that do make moral decisions, but they do so by following constraints specified by a human operator. Even if fully autonomous such systems can only "recognize" and react to a morally sensitive situation if explicitly programmed to do so [12]. For example, Apple's Siri would not give answers to explicit requests for suicide methods, but cannot handle implicit "calls for help"[2]. Explicit and functional moral agents have the capacity to use their autonomy to recognize and act upon a morally sensitive situation, advancing beyond their explicit programming. Machine ethics is particularly concerned with the behavior of implicit and explicit agents.

4. Certifying the behavior of moral agents

How to implement moral artificial agents, namely the implicitly and explicitly moral agents, remains a challenge for both researchers and engineers. However, implementation alone is not sufficient. A moral agent, or a machine such an agent controls, needs to be certified before it is allowed to impact society. Certification informs consumers and experts of the properties of a product, a system, or a person in a position of responsibility. Each of the different types of moral agents should be subject to a different process of certification, just as the process of certifying surgeons is different than the process of certifying toasters.

In the case of ethical impact agents, it is not the agent itself but how this agent is used that has a moral impact on society. Explainability, transparency and accountability of AI particularly applies to this class of moral agents and could be used in the process of their certification.

It has been argued that formal verification can play an important role in certifying implicitly ethical agents [8]. Formal verification is the process of using formal methods of mathematics to prove or refute the correctness of intended algorithms underlying a system with respect to a certain property. The challenge in verification is to formally specify the property of ethical behavior.

With explicit ethical agents it is least clear what the process of certification should entail. If the agent is adaptive, then we need to show that not only does the agent behave ethically at a given moment, but also that the agent will not adapt towards unethical behavior. It has been proposed that agents can be tested for ethical behavior by using a moral Turing test, a version of the Turing test involving asking questions of morality to the agent and comparing the answers to those obtained from a moral expert [2,3]. It has also been argued that the moral Turing test is not sufficient [5].

The question of how to make autonomous yet moral machines remains widely open from philosophical and AI research aspects, however it is also a legal problem [7]. De-

[2]https://www.cnbc.com/2018/06/06/siri-alexa-google-assistant-responses-to-suicidal-tendencies.html

ciding who decides which morality should an autonomous system implement and who should such a system be explainable or accountable to, remains a wider societal issue.

References

[1] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[2] C. Allen, G. Varner, and J. Zinser. Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical AI*, 12:251–261, 2000.

[3] M. Anderson and S. Leigh Anderson. Geneth: A general ethical dilemma analyzer, 2014.

[4] S. Anderson, M.and Leigh Anderson. Machine ethics: Creating an ethical intelligent agent. *AI Magazine*, 28(4):15, 2007.

[5] T. Arnold and M. Scheutz. Against the moral turing test: accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology*, 18(2):103–115, Jun 2016.

[6] T. Bolukbasi, K.W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4356–4364, USA, 2016. Curran Associates Inc.

[7] J. Bryson and A.F.T. Winfield. Standardizing ethical design for artificial intelligence and autonomous systems. *IEEE Computer*, 50(5):116–119, 2017.

[8] L. A. Dennis, M. Fisher, M. Slavkovik, and M. P. Webster. Formal Verification of Ethical Choices in Autonomous Systems. *Robotics and Autonomous Systems*, 77:1–14, 2016.

[9] L. A. Dennis and M. Slavkovik. Machines that know right and cannot do wrong: The theory and practice of machine ethics. *IEEE Intelligent Informatics Bulletin*, 19(1), 2018.

[10] V. Dignum. Responsible autonomy. In *Proceedings of the 26th IIJCAI*, pages 4698–4704, 2017.

[11] V. Dignum. Accountability, responsibility, transparency - the ART of AI. In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence, ICAART 2018, Volume 1, Funchal, Madeira, Portugal, January 16-18, 2018.*, page 7, 2018.

[12] S. Dyrkolbotn, T. Pedersen, and M. Slavkovik. On the distinction between implicit and explicit ethical agency. In *AAAI/ACM AIES conference*, New Orleans, USA, 2018.

[13] A. Etzioni and O. Etzioni. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, pages 1–16, 2017.

[14] A. G. Ferguson. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. NYU Press, 2017.

[15] B. Gert and J. Gert. The definition of morality. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition, 2017.

[16] J. Lau, B. Zimmerman, and F. Schaub. Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of ACM Human-Computer Interaction*, 2(CSCW):102:1–102:31, November 2018.

[17] H. Lipson and M. Kurman. *Driverless: Intelligent Cars and the Road Ahead*. The MIT Press, 2017.

[18] D. Maevsky, A. Bojko, E. Maevskaya, O. Vinakov, and L. Shapa. Internet of things: Hierarhy of smart systems. In *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 2, pages 821–827, 2017.

[19] J. H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, July 2006.

[20] S.Y. Nof. *Handbook of Industrial Robotics*. Number v. 1 in Electrical and electronic engineering. Wiley, 1999.

[21] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty. Stakeholders in explainable AI. *CoRR*, abs/1810.00184, 2018.

[22] S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 3 edition, 2015.

[23] N. Sambasivan and J. Holbrook. Toward responsible AI for the next billion users. *Interactions*, 26(1):68–71, 2019.

[24] H. Simonis. Constraints in computational logics. chapter Building Industrial Applications with Constraint Programming, pages 271–309. Springer-Verlag New York, Inc., 2001.

[25] I. Ulrich, F. Mondada, and J.-D. Nicoud. Autonomous vacuum cleaner. *Robotics and Autonomous Systems*, 19(3):233 – 245, 1997. Intelligent Robotic Systems SIRS'95.

[26] S. Wachter, B. Mittelstadt, and L. Floridi. Transparent, explainable, and accountable ai for robotics. *Science Robotics*, 2(6), 2017.

[27] W. Wallach and C. Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2008.

[28] W. Wallach, C. Allen, and I. Smit. Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI Society*, 22(4):565–582, 2008.

[29] F. M. Zanzotto. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243 – 252, 2019.

[30] C. Zopounidis. Multicriteria decision aid in financial management. *European Journal of Operational Research*, 119(2):404 – 415, 1999.