Basic Issues in AI Policy

Vincent C. Müller[1,2(✉)]

[1] Ethics of Technology, Philosophy and Ethics Group, Faculty of IE/IS, Technical University of Eindhoven (TU/E), Atlas 9.329, De Zaale, P.O. Box 513, 5600 Eindhoven, MB, The Netherlands

[2] IDEA Centre, Faculty of Philosophy, Religion and History of Science (PRHS), University of Leeds, Leeds, UK

v.c.muller@tue.nl

http://www.sophia.de/

Abstract. This extended abstract summarises some of the basic points of AI ethics and policy as they present themselves now. We explain the notion of AI, the main ethical issues in AI and the main policy aims and means.

1 AI

The term "Artificial Intelligence" (AI) is now used in two main meanings:

(a) AI is a research programme to create computer-based agents that can show complex behaviour, capable of reaching goals [1], OR
(b) AI is a set of methods employed in the AI research programme for perception, modelling, planning, and action: machine learning (supervised, reinforced, unsupervised), search, logic programming, probabilistic reasoning, expert systems, optimisation, control engineering in robotics, neuromorphic engineering, etc. Many of these methods are also employed outside the AI research programme [2–5]

The original research programme (a) from the "Dartmouth Conference" in 1956 onwards was closely connected to the idea that computational models can be developed for cognitive science of natural intelligence and then implemented on different hardware, i.e. on computing machines. This programme ran into various problems in the "AI Winter" ca. 1975–1995 and the word "AI" got a bad reputation; it thus branched into several technical programmes that used their own names (pattern recognition, data mining, decision support system, data analytics, cognitive systems). After 2000, AI saw a resurgence, with faster hardware, more data, and a stress on neural network machine-learning systems. From ca. 2010 "AI" became a buzzword that resonated in circles outside computer science; now everyone wants to be associated with AI. As a result, the meaning of "AI" is currently broadening towards (b). (More details on this development and the contrast to robotics in [6].

2 AI Ethics

As a result of the popularity of AI, around 2015, the possible negative effects of AI on humanity came into focus as well, e.g. whether AI would take human jobs, decide over

Grau-Ruiz, Maria A.(ed.): Interactive robotics: Legal, ethical, social and economic aspects. 3–9, 2022.
https://doi.org/10.1007/978-3-031-04305-5_1

human lives, spell the end of privacy, kill humans with autonomous weapons or just take over the world entirely in a 'singularity'. Eurobarometer surveys showed that European citizens' views on robots had become more negative [7, 8].

Work on *ethics of AI* had existed since the 1950ies but it was a very small minority of researchers who worked on it, mostly in the context of *computer ethics and data ethics* (esp. on privacy) [for a history, see 9]. AI ethics is closely connected to the philosophy of AI. As a field, it is reaching maturity in the present years, and surveys of the state of the art have only appeared since 2020 [6, 10–12].

AI ethics is an application of the methods of ethics, i.e. it is a branch of "applied ethics". Ethics is the classical discipline of philosophy that describes how to evaluate actions as right or wrong, i.e., how to use normative statements that involve values. The basis for such ethical judgment is typically an evaluation of personal virtues (positive character traits, e.g. honesty or friendliness), of ethical rules or duties (e.g. "do not steal") and of the consequences of an action (typically in terms of utility). Ethical *dilemmas* occur when rules conflict (e.g. honesty vs. politeness) or when rules conflict with outcomes (e.g. honesty results in harm for an innocent person). It is clear that a comprehensive theory of ethics, and thus a comprehensive applied ethics, must take into account all three components: the person performing the action, the type of action and the consequences of the action.

Sometimes it has been assumed that if machines act in ethically relevant ways, then we need a machine ethics, for "ethical machines", for machines as subjects, rather than for the human use of machines as objects [13–16]. However, this is very likely a misunderstanding since there is general agreement that current and foreseeable AI systems do not have what it takes to be responsible for their actions (moral agents), or to be systems that humans should have responsibility towards (moral patients) [17]. So, the responsibility remains firmly with the humans and for the humans – as well as other animals.

What is at stake in AI ethics is the human design and use of AI. There seems to be reasonable agreement that the main ethical issues are:

- Privacy & Surveillance
- Manipulation of Behaviour
- Opacity of AI Systems
- Bias in Decision Systems
- Human-Robot Interaction
- Automation and Employment
- Autonomous Systems and Responsibility
- Machine Ethics
- Artificial Moral Agents
- Singularity

For details on each of these, see the 10 sections on the "main debates" in [6].

If successful, AI ethics says what is right and what is wrong, i.e. it defines what "ethical AI" would be like (see below). Knowing what ethical AI is like is a significant input to *AI policy* (see below) because it defines what the main policy *aims* of AI policy should be.

3 Ethical AI

3.1 Explanation

If AI is designed and used ethically, as defined by AI ethics, then it is *ethical AI*. This desirable state has often been expressed in different terms that sound less academic, e.g. using the term "Human-centric AI" is an approach to AI that judges AI by its positive effects on humans, which captures most of what we want from ethical AI. It stresses that AI should serve human well-being, guarantee human rights or dignity, and have positive effects on society. All of these things are traditionally captured in ethics, which takes into account rights and positive consequences (in deontological and consequentialist considerations).

Some years ago, there was talk of developing "trust in AI", especially after the Eurobarometer had shown that there was little trust – focusing on threats to jobs from AI ? and by autonomous robots. However, one can trust a system for good and for bad reasons, so what matters is whether it *deserves* to be trusted, whether it is "trustworthy". Thus "trustworthy AI" became the central term in the EC White paper [18] and the slogan for EC AI policy.

Much speaks in favour of just calling this "good AI", but "ethical AI", "human-centred AI" or "trustworthy AI" will do just as well.

3.2 AI Ethics Guidelines

In the effort to explain what "ethical AI" might mean, in ca. 2015–2020, many institutions published "ethics guidelines" (under several terms) and to explain how they are in favour of ethical AI. There is a map on https://aiethicslab.com/big-picture/, a list on https://inventory.algorithmwatch.org (and http://www.ptai.org/TG-ELS/policy) and there are recent surveys of these [19, 20]. What these guidelines do is usually not very exciting; they look at the points where ethical issue do occur, like the list of 10 above, and say that we should be ethical – rarely going into the points where there is actually a difficulty in saying *what is ethical*, i.e. a dilemma. For example, the Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission [21] says:

– AI should guarantee:
– Human agency and oversight
– Technical robustness and safety
– Privacy and data governance
– Transparency
– Diversity, non-discrimination and fairness
– Societal and environmental wellbeing
– Accountability

This really just says that AI must guarantee responsibility and liability (1 & 7), positive utility in the long run (2 & 6) and safeguard rights like privacy, autonomy and justice (3 & 4 & 5). None of this comes as news to AI ethics.

The question is *what should be done* to make sure ethical AI is a reality. This is where AI policy comes in.

4 AI Policy

4.1 Definition

A policy is a set of principles to guide organisational actions (typically of states or supranational organisations). A policy consists of two main components: policy *aims* and policy *means* to further those aims.

4.2 Policy Aims

In our case of AI, policy *aims* are constituted by a) ethical principles on AI, plus b) general policy aims, such as those formulated in the 17 UN Sustainable Development Goals (https://sdgs.un.org).

It is thus important for AI policy to specify the specific (a) and general (b) aims it pursues.

While there is significant agreement on the specific ethical principles on AI, general policy aims will often differ between nation states, e.g. the US has a high sensitivity to bias and discrimination, China and the US place great emphasis on geostrategic aims (will thus be reluctant to limit automated weapons), the EU will be sensitive to monopolies and places great emphasis on privacy, many states will put a higher emphasis on economic development than on justice, etc.

Which policy aims will *actually be pursued* also depends on powers that the state actors are subjected to, e.g. public opinion, lobbying, technical feasibility, cost, etc.

Setting policy aims, e.g. through AI ethics guidelines, should not be identified with policy, since policy also involves practical activity to achieve these aims.

4.3 Policy Means

Policy *means* are the practical instruments and methods with which we can further the policy aims. It appears that the main bottlenecks of AI policy, at the moment, is with the appropriate policy means.

The first thought of policy aims goes for legal regulation, but it is important to remember the broad range of other options that exist:

- educational efforts (e.g. curriculum of AI degrees)[1]
- framework for legal liability (e.g. insurance)
- impact assessment tools
- legal regulation
- PR measures
- public spending
- self-assessment frameworks
- self-regulation (in industry)
- self-regulation (e.g. a “Hippocratic oath”)
- supporting ethics by design
- taxation
- technical standards (in a framework of legal regulation)

There is material that can help in this area and in politics and political science there is ample experience in developing policy & in facing the bottlenecks of practical policy means.

From other cases of ethics-driven policy, there are models that we can follow, e.g. medical ethics [22] or engineering ethics [23], or professional ethics in various areas. What can be learned from traditional computer ethics, esp. data ethics [24, 25]. What are the ways forward in a professional ethics for AI engineers [26]? Are the principles we have "actionable" [cf. 27]? Perhaps self-assessment is the way, such as the ALTAI self-assessment tool (EC, Unit A1; July 2020): https://altai.insight-centre.org [28]. Further to "ethical AI", there is also a movement to use AI to further other social aims, i.e. to use "AI for social good" [29] – this goes beyond AI policy.

4.4 EU Regulatory Efforts

In the EU, a dominant notion for policy means is "risk". The AI HLEG says there are two kinds of risks [21]: Risks for fundamental rights, including personal data and privacy protection and non-discrimination. Risks for safety and the effective functioning of the liability regime. These risks are then classified in terms of high-risk vs. low-risk. (This is a bit misleading since the rights are not at risk, the risk is that the rights could be violated. Typically, one distinguishes *rights* (that exist independently of positive or negative consequences) from *risks* (the probability of a negative consequence). Any ethics will involve both a notion of rights/obligations and a consideration of outcomes.)

This is combined with moves towards certification, ex ante and post hoc, voluntary and involuntary, self-certification and certification through official bodies. (In Germany through the BSI, DIN and VDA [cf. 30]).

Acknowledgments. This work was sponsored by the European Commission under the INBOTS and IA-AI projects.

References

1. McCarthy, J., Minsky, M., Rochester, N., Shannon, C.E.: A proposal for the Dartmouth summer research project on artificial intelligence. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html. Accessed October 2006
2. Russell, S.: Human Compatible: Artificial Intelligence and the Problem of Control. Viking, New York (2019)
3. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 4th edn. Prentice Hall, Upper Saddle River (2020)
4. Görz, G., Schmid, U., Braun, T.: Handbuch der künstlichen Intelligenz, 5th ed. De Gruyter, Berlin, p. 976 (2020). (in d)
5. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic Books, New York (2018)
6. Müller, V.C.: Ethics of artificial intelligence and robotics. In: Zalta, E.N. (ed.) Stanford Encyclopedia of Philosophy, vol. Summer 2020, pp. 1–70. Palo Alto: CSLI, Stanford University (2020)

7. Eurobarometer: Public attitudes towards robots, vol. 382. http://ec.europa.eu/public_opinion/archives/ebs/ebs_382_sum_en.pdf
8. Eurobarometer: Autonomous systems, vol. 427. https://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_427_en.pdf
9. Müller, V.C.: History of Digital Ethics. In: Véliz, C. (ed.) Oxford Handbook of Digital Ethics. Oxford University Press, Oxford. forthcoming
10. Coeckelbergh, M.: AI Ethics. MIT Press, Cambridge, Mass (2020)
11. Gibert, M.: Faire la moral aux robots: Une introduction à l'éthique des algorithmes. Montréal: Atelier 10 (2020)
12. Gordon, J.-S., Nyholm, S.: Ethics of Artificial Intelligence, Internet Encyclopedia of Philosophy. https://iep.utm.edu/ethic-ai/
13. Floridi, L., Saunders, J.W.: On the morality of artificial agents. Minds Mach. **14**, 349–379 (2004)
14. Moor, J.H.: The nature, importance, and difficulty of machine ethics. IEEE Intell. Syst. **21**(4), 18–21 (2006). https://www.researchgate.net/publication/220629129_The_Nature_Importance_and_Difficulty_of_Machine_Ethics
15. Anderson, M., Anderson, S.L. (eds.): Machine Ethics. Cambridge University Press, Cambridge (2011)
16. Wallach, W., Asaro, P.M. (eds.): Machine Ethics and Robot Ethics. Routledge, London (2017)
17. Müller, V.C.: Is it time for robot rights? Moral status in artificial entities. Ethics Inf. Technol. 1–9, forthcoming. https://www.springer.com/journal/10676
18. European Commission: White paper on artificial intelligence: A European approach to excellence and trust, vol. COM (2020) 65, no. 19.2.2020, pp. 1–27 (2020). https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en
19. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020).https://doi.org/10.1007/s11023-020-09517-8
20. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
21. AI HLEG: High-level expert group on artificial intelligence: ethics guidelines for trustworthy AI, European Commission, pp. 1–37 08 April 2017. https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence
22. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. **1**(11), 501–507 (2019). https://doi.org/10.1038/s42256-019-0114-4
23. Van de Poel, I.R., Royakkers, L.M.: Ethics, Technology and Engineering. Wiley-Blackwell, Oxford (2011)
24. Mittelstadt, B.D., Floridi, L.: The ethics of big data: current and foreseeable issues in biomedical contexts. Sci. Eng. Ethics **22**(2), 303–341 (2016). https://doi.org/10.1007/s11948-015-9652-2
25. Floridi, L., Taddeo, M.: What is data ethics? Phil. Trans. R. Soc. A **374**(2083) (2016)
26. Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mokander, J., Floridi, L.: Ethics as a service: a pragmatic operationalisation of AI Ethics, 11 Feburary 2021. arXiv:2102.09364v1. https://arxiv.org/abs/2102.09364
27. Stix, C.: Actionable principles for artificial intelligence policy: three pathways. Sci. Eng. Ethics **27**(1) 1–17 (2021). https://doi.org/10.1007/s11948-020-00277-3
28. HLEG: Assessment List for Trustworthy AI (ALTAI) for self- assessment no. 17 June 2020. https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence

29. Floridi, L., Cowls, J., King, T.C., Taddeo, M.: How to design AI for social good: seven essential factors. Sci. Eng. Ethics **26**(3), 1771–1796 (2020). https://doi.org/10.1007/s11948-020-00213-5
30. Wahlster, W., Winterhalter, C.: Deutsche Normungsroadmap Künstliche Intelligenz, p. 232. DIN/DKE, Berlin (2020)