Technology and Regulation

Ethics of artificial intelligence: Some ethical issues and regulatory challenges

Mark Coeckelbergh*

responsibility, privacy, transparency, explainability, bias, AI policy, ethics by design

This article offers a brief overview of some of the ethical challenges raised by artificial intelligence (AI), in particular machine learning and data science, and summarizes and discusses a number of challenges for near-future regulation in this area. This includes the difficulties of moving from principles to more concrete measures and problems with implementing ethics by design and responsible innovation.

mark.coeckelbergh@univie.ac.at

1. Introduction

AI is already having a pervasive impact today as it is embedded in everyday digital technological systems, and its promises and attractions are likely to increase this impact in the near future. It is likely to have impact in many domains such as transport, marketing, health care, finance, security, science, education, entertainment, agriculture, and manufacturing.

While AI is likely to have many benefits, it also raises a number of ethical issues, some of which are well-known (e.g., privacy) and others which have to do with specific technologies and applications, such as bias created by machine learning and the related data science, or responsibility attribution problems that emerge from these methods and processes. Many of these issues do not only play out at an individual level, but also concern transformations in societies and economies. This is especially the case with AI-powered automation, which enables machines to take over tasks from humans.

This article gives a brief overview of some of the ethical issues and summarizes and discusses a number of challenges for near-future regulation in this area. The focus is on artificial intelligence applications that involve machine learning and data science.

2. Some ethical issues raised by artificial intelligence

Since AI and especially machine learning methods involve a process of data collection, processing, and sharing, a first issue – shared with many other digital technologies – concerns the question whether the privacy of individuals is respected and even whether they know that their data is collected at all. In the context of AI and data science these questions are especially urgent since often users do not know that AI is behind an application they use (e.g., an app on their phone) and since often data given in one context and one domain are then used by another party in another context and another domain, without the knowledge and consent of the people who gave their data.

Another well-known problem is data security: all these systems are networked and may be hacked for malicious purposes (e.g., cybercrime, cyberwar). The technology also relies on vulnerable material infrastructures: AI and other information systems are not entirely made of immaterial code but are embedded in material technological systems and require material infrastructures, which can be disrupted or destroyed.

Moreover, a problem that becomes especially relevant in the case of AI is attribution of responsibility. Since technologies cannot be responsible moral agents and are hence a-responsible, the only way to ensure responsible action is to make humans responsible. However, in technological action it is notoriously difficult to ascribe moral responsibility due to the so-called problem of "many hands": many people are involved in the often long causal histories that lead to a particular outcome. If there is a problem with the end result, say a recommendation, it is difficult to figure out who was responsible. And since AI is often part of a larger technological system and data histories, it is difficult to figure out if "the AI" caused the problem or some other part of the system. There are not only many hands but also many things.

Responsibility is especially problematic when people who use the systems are lured by the potential of the technology and use it without much hesitation, but are ignorant about most of the system and its history, for example how the data has been generated and combined. People using the systems are supposed to take responsibility but this becomes difficult if they don't know what they are doing.

* Prof. Dr. Mark Coeckelbergh is a full Professor of Philosophy of Media and Technology at the Department of Philosophy of the University of Vienna and the President of the Society for Philosophy and Technology (SPT).

Mark Coeckelbergh, Ethics of artificial intelligence: Some ethical issues and regulatory challenges, Technology and Regulation, 2019, 31–34 • https://doi.org/10.26116/techreg.2019.003 • ISSN: 2666-139X

But even experts don't always know everything, and this leads us to the problem of transparency and explainability. It is not always clear what is happening in the process, and this is especially the case for so called "black box" systems like machine learning that uses neural networks where technically an outcome (recommendation) cannot be traced back to a chain of decisions or reasoning as in decision tree models. Such systems are thus opaque. This is an ethical problem since people should have the right to know why a decision that affects them was taken. If a decision cannot be explained, this is unjust. Explainability is thus a moral requirement.

The problem of bias, furthermore, is especially challenging. Bias means that some individuals or groups are disadvantaged by the outcome of the system. Although problems of bias and discrimination have always been present in societies and cultures, the concern here is that the AI technology may perpetuate these and increase their impact. Bias is often unintended, but may be generated at various stages of the machine learning and data science process. Bias can arise in the selection of the data set, in the training dataset itself, in the algorithms used, in the application dataset, and indeed in wider society. Consider for example an AI that is trained on text data from the internet, which contains bias in the particular texts or even in the language (e.g., English). Perhaps bias cannot be avoided, in the sense that surely algorithms used for making decisions (e.g., about job applicants) are used for discriminating (e.g., between suitable candidates and others). But the question is always if a particular bias and discrimination is unjust and unfair. An answer to that question is not a merely technical question but an ethical and political one; it depends on our views of justice and on what kind of society we want.

Finally, in so far as AI is used for automation it also impacts work and the future of society. Many authors warn of unemployment and raise the question if a re-structuring of our social institutions is necessary (e.g., basic economy) to answer some of these challenges. This also makes us think about the political question *who* will decide about the technological future.

3. Addressing ethics of AI issues

While many policy makers that seek regulation of AI agree that something needs to be done in response to these ethical problems, they face a number of challenges. For a start, they need to answer the following questions: they need to figure out *what* should be done, justify *why* it should be done, *by whom* it should be done, and so on. For example, it is not easy to deal with the problem of bias: it is not clear what, exactly, should be done to avoid it as much as possible, and who should take action. And if existing regulation is seen as insufficient, new regulation should be justified: why is it needed, why is the existing regulation not enough? For example, in the case of data protection and privacy but also with regard to transparency and explainability, some argue that the European General Data Protection Regulation (GDPR)[1] instrument, which provides enforceable legislation, is sufficient; others argue that it does not provide enough protection against the risks of automated decision-making when it comes to explainability: there is only a right to information but this does not require full explainability.[2]

4. Guidance

So what about the future? The past year has seen a large number of policy documents that address ethics of AI, both from the public and the private sector. For example, already under the Obama presidency, the U.S. government published a report on the future of artificial intelligence[3] and last year many European countries published reports and strategies, for example the UK[4] and France[5]. Many documents propose trustworthy AI and explainable AI, and this has been reflected in supranational work on AI policy. In April 2018, the EU set up a new High-Level Expert Group on AI (HLEG AI) which has recently produced a document with ethical guidelines for AI (European Commission 2019). Earlier the European Group on Ethics in Science and New Technologies (EGE)[6] released a statement on AI which also proposes a number of principles. China and other major global players also have an AI strategy that includes ethics. For example, China has a development plan that recommends minimizing risk.[7] In addition, there also have been civil society actors commenting or campaigning with regard to AI, for example to ban autonomous weapons or to protect the privacy of citizens. And the Institute of Electrical and Electronics Engineers (IEEE), a large international technical professional organization, has taken a Global Initiative on Ethics of Autonomous and Intelligent Systems which has resulted in guidelines for ethical design.[8] And companies such as Google also published AI principles. They are not necessarily opposed to regulation; Apple's CEO Tim Cook has said that tech regulation is inevitable.[9] However, most industry players seem to prefer a minimal degree of regulation. This is a challenge for those who wish to move towards more substantial regulatory efforts.

Most policy proposals concerning AI ethics start from a number of ethical principles. For example, the HLEG AI starts from fundamental rights (human dignity, freedom of the individual, respect for democracy, justice and the rule of law, and citizens' rights) and a number of ethical principles, some of which are known from bioethics (the no harm principle, for example) but also explicability. These principles are relevant to AI in the form of machine learning: no harm requires that AI algorithms avoid discrimination, manipulation, and negative profiling, and explicability is interpreted as requiring that AI systems be auditable and comprehensible.[10]

However, this approach in terms of principles raises a number of challenges.

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation 2016 [OJ L 119, 4.5.2016, pp. 1–88].

[2] Sandra Wachter, Brent Mittelstadt and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' *International Data Privacy Law*, Volume 7, Issue 2, 1 May 2017, 76–99,

[3] National Science and Technology Council Committee on Technology, 'Preparing For the Future of Artificial Intelligence' (Executive Office of the President, Office of Science and Technology Policy (OSTP) 2016).

[4] House of Commons, 'Algorithms in Decision-Making, Fourth Report of Session 2017-19' (2018).

[5] Cédric Villani, 'For a Meaningful Artificial Intelligence - Towards a French and European Strategy' (2018) https://www.aiforhumanity.fr/pdfs/Mission-Villani_Report_ENG-VF.pdf.

[6] European Group on Ethics in Science and New Technologies (EGE), 'Statement on Artificial Intelligence, Robotics and "Autonomous" Systems' (European Commission, Directorate-General for Research and Innovation 2018).

[7] 'New Generation Artificial Intelligence Development Plan. (新一代人工智能发展规划) Translation Available at https://Flia.Org/Notice-State-Council-Issuing-New-Generation-Artificial-Intelligence-Development-Plan/ (State Council of China 2017).

[8] https://ethicsinaction.ieee.org/

[9] https://www.businessinsider.de/apple-ceo-tim-cook-on-privacy-the-free-market-is-not-working-regulations-2018-11

[10] European Commission High-Level Expert Group on Artificial Intelligence (HLEG), 'Ethics Guidelines for Trustworthy AI' (European Commission 2019) https://ec.europa.eu/futurium/en/ai-alliance-consultation.

5. Challenges ahead

First, it is not clear if these expressions of concern for ethics of AI will actually lead to concrete regulation (when it comes to public actors) or concrete actions by corporations (private sector). While, for example, the European Commission set up procedures to stimulate uptake by stakeholders, there is no guarantee that this will actually happen. There is a risk that ethics are used as a fig leaf that helps to ensure acceptability of the technology and economic gain but has no significant consequences for the development and use of the technologies.

Second, even if stakeholders intend to do something with these documents, it is a challenge for regulation to move from more or less vague and abstract principles to more concrete methods, procedures, laws, and institutions. What are the concrete outcomes? Will there be new directives? New laws? Will there be a new agency that can monitor the implementation? While the European Commission document goes some way towards operationalization (further than any of the other documents I read), there is still a lot of work needed in this direction. It remains a huge challenge to bridge between abstract principles and concrete practices.

Third, one of these practices includes development and design of technologies; hence one may propose an ethics by design approach and similar measures. A proactive approach to technology ethics requires that ethics does not only come afterwards, by means of regulation after the technology is already developed, but that a regulatory framework is created to stimulate and (hopefully) ensure that ethics is already taken into account in earlier stages: in the development of the technology. For example, ethics by design could mean that it is required that traceability is ensured at all stages.[11] It is a challenge to think about how to technically implement ethics. For example, Winfield et al.[12] have called for implementing an 'ethical black box' in robots and autonomous systems which records data from sensors and the internal system; this could also be applied to AI. More generally, it is challenging to think about how to ensure explainability in technical ways. A related idea is responsible innovation[13], which requires that all kinds of stakeholders are involved in these earlier stages of development, potentially rendering the whole process more democratic and just.

These ideas also support the vision that regulation need not all be about banning things. We need a positive and constructive ethics of AI, which is not only about regulation in the sense of constraints but which also concerns the question of the good life and human and societal flourishing. Before thinking about concrete regulation, policy makers are challenged to develop a positive vision about where AI should take us.

6. Ethics by design and responsible Innovation

However, ideas such as ethics by design and responsible innovation and their implementation have their own barriers. It may be difficult to operationalize the general principles.

A. First, it is already great that explainability is operationalized as traceability in the HLEG guidelines, but what exactly does traceability mean? To find out what exactly should be done is itself a research question.
B. Second, ethics by design sounds great but it is not so easy to foresee the unintended consequences of new technologies at an early stage. More thinking needs to be done about concrete methodologies and techniques.
C. Third, it is hard to see how responsible innovation can really be implemented when there is a concentration of power in the hands of a relatively limited number of powerful actors, including a small number of large corporations: it seems that a handful of companies decide the future of AI.[14]
D. Fourth, can we really make fully explicit our values[15], given that ethical knowledge is partly tacit?
E. Finally, ethics by design, value sensitive design, responsible innovation, etc. work on the assumption that the technology will be developed[16]; is there also at least the possibility that the technology or the applications can be halted? How much room is there for deciding otherwise?

Fourth, there needs to be more interaction between legal and ethical expertise. For example, there are interesting questions with regard to which legal instruments can and should be used for dealing with problems of responsibility. For instance, whereas criminal law requires the intention to do harm, negligence asks the question whether a person was under a duty of care to prevent harm; this seems more applicable to AI and the people involved in its processes. Product liability, furthermore, does look at the fault of the person but has the company who produced the AI pay for damages, regardless of fault.[17] This could also be an interesting route to deal with responsibility issues. More generally, there needs to be a discussion about which legal instruments (existing or new) can and should deal with the ethical problems indicated.

Fifth, more generally, there is still a gap in understanding between people coming from the humanities and social sciences and those who have a technical background. A lack of interdisciplinarity can hinder the effectiveness of policy making in an area such as ethics of AI, when parties involved are constrained by their disciplinary understandings. Similarly, transdisciplinarity is needed in the sense that experts from academia need to reach out to (other) stakeholders and vice versa. We need to think about ways to bring together people and domains of knowledge and experience, not only in policy-making and professional life but also in the stage of education.

Sixth, it seems that given the nature of the technology, the problems are global and need to be addressed at a global level. But this is difficult when policy-making is largely happening at nation state level. How effective is it to take regulatory measures at the national level when the technology is developed and used across borders?

Finally, AI ethics policy is also a matter of priorities. There may be other technologies that also stand in need of regulation. And there may be national and global issues that also require our ethical and

[11] Virginia Dignum and others, 'Ethics by Design: Necessity or Curse?', *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18* (ACM Press 2018) http://dl.acm.org/citation.cfm?doid=3278721.3278745 accessed 1 May 2019.

[12] Alan FT Winfield and Marina Jirotka, 'The Case for an Ethical Black Box' in Yang Gao and others (eds), Towards Autonomous Robotic Systems (Springer International Publishing 2017)

[13] René von Schomberg, 'Towards Responsible Research and Innovation in the Information and Communication Technologies and Security Technologies Fields' (European Commission 2011).

[14] Paul Nemitz, 'Constitutional Democracy and Technology in the Age of Artificial Intelligence' DOI 10.1098/RSTA.2018.0089 - Royal Society Philosophical Transactions A

[15] Paula Boddington, Towards a Code of Ethics for Artificial Intelligence (1st ed. 2017 edition, Springer 2017).

[16] Kate Crawford and Ryan Calo, 'There Is a Blind Spot in AI Research' (2016) 538 Nature 311.

[17] Jacob Turner, Robot Rules - Regulating Artificial Intelligence (Palgrave Macmillan 2019).

political attention, such as social-economic injustices and climate change. A good AI policy that aims to be ethical needs to address this question of priorities, which is an ethical and political question.

If these barriers can be overcome, there is a chance for effective and good regulation of AI in an ethical direction and, more generally, an AI future that we want.