Establishing the rules for building trustworthy AI

The European Commission's report 'Ethics guidelines for trustworthy AI' provides a clear benchmark to evaluate the responsible development of AI systems, and facilitates international support for AI solutions that are good for humanity and the environment, says Luciano Floridi.

Luciano Floridi

AI is revolutionizing everyone's life, and it is crucial that it does so in the right way. AI's profound and far-reaching potential for transformation concerns the engineering of systems that have some degree of autonomous agency. This is epochal and requires establishing a new, ethical balance between human and artificial autonomy.

Careful planning rather than beta testing

As a new kind of autonomous, smart agency, AI could bring enormous benefits — individually, socially and environmentally. It could represent a force for good in a world that is increasingly complex and requires sophisticated solutions to deal with large-scale and interrelated issues. The 17 UN Sustainable Development Goals show that humanity is struggling with many challenges, on many vital fronts, and it would be unwise not to make use of AI solutions. However, what processes and decisions are going to be delegated to AI systems, what kinds of effects the trade-offs between human and artificial agency are going to have, and what forms of assessment, control, revision and redressing must be put in place, are crucial questions that should not be answered through trial and error. AI should never be beta-tested on humans or the environment. The development of AI requires socio-political deliberation and consensus, in view of a long-term strategy about what kind of AI should be developed, for what purpose, for whom, and according to which ethical priorities. This is a main aim of the ethics guidelines report from the European Commission (EC).

The report, published on 8 April 2019 after several versions and more than 500 public consultations, is put together by an independent, High-Level Expert Group (HLEG)[1]. The HLEG was appointed by the EC in June 2018 and consists of 52 experts (disclosure: I am one of them), with relevant expertise from academia, civil society and industry. The work of the HLEG is expected to inform the European Union's (EU) policies and legislation about AI, to support the implementation of the EU strategy on AI, and to serve as the steering group for the European AI Alliance's work.

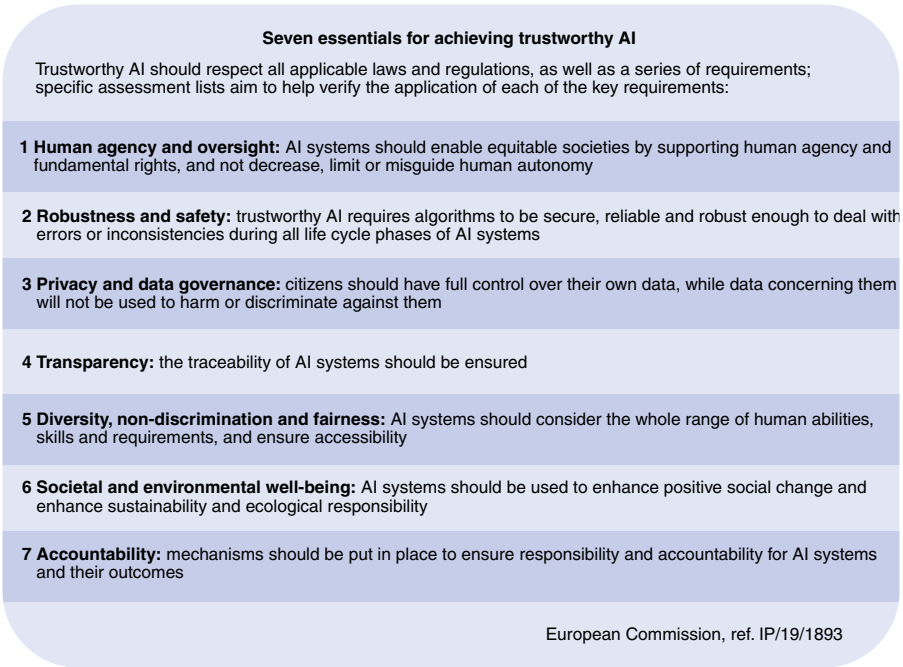

Fig. 1 | The seven ethical principles grounding the EU 'Ethics guidelines for trustworthy AI'. Adapted from ref. [12], European Commission.

The guidelines support a responsible approach to the development of AI, which should be (1) lawful, respecting all applicable laws and regulations; (2) ethical, respecting ethical principles and values (Fig. 1 summarizes the principles grounding the guidelines, which were informed[2] by the AI4People's research[3]); and (3) robust, both technically and in terms of its social environment.

Since AI will become increasingly important and pervasive, it must work reliably, in ways that anyone can trust will be for the benefit of humanity and the whole environment. The alternative is that AI may be misused, overused or underused[3]. Ethical uncertainty breeds both reckless risk-taking and excessive caution. This is why the guidelines are so important. They represent a good step in the right direction of a clear, shared and socially preferable framework for ethical AI.

Ethics first to inform legislation

The guidelines have been praised and welcomed by many, but have also been criticized[4,5] for being weak, because they are part of a mere self-regulatory strategy, which is not legally enforced, and unhelpful, because they are too general, and join so many other initiatives that have so far had little impact. These and similar criticisms can be countered.

First, the guidelines contain principles and clarifications that are robust, in terms of social expectations, and consistent with the current state of the debate on the ethics of AI. Of course, both law and ethics about AI are needed. The guidelines presuppose and are aligned with the EU legislation. The EU is at the forefront of the international debate on AI, also thanks to the General Data Protection Regulation (GDPR). Ethics can contribute to the shaping of new legislation (for example, about facial recognition systems) or act as a guide in its absence.

Sometimes, ethics is needed to interpret existing legislation (for example, the GDPR). Other times, ethics may recommend not to do something that legislation does not prohibit (for example, leaving a medical decision entirely to an algorithm without supervision or explanation), or recommend to do something that legislation does not require (for example, designing an algorithm that minimizes the environmental impact of domestic central heating). In all these cases, compliance with the law is necessary but insufficient, and, as the guidelines acknowledge, it must be complemented by a post-compliance 'soft ethics' approach[6], because the law provides the rules of the game, but does not indicate how to play well according to the rules.

Second, granted: the guidelines are not very original or innovative, but that would have been astonishing and perhaps a bit concerning, after more than a half a century of discussion on the topic[7,8]. There are in fact currently more than 70 frameworks and lists of principles about the ethics of AI[9,10]. This mushrooming of declarations is generating inconsistency and confusion, among stakeholders, regarding which document may be preferable. It also puts pressure on private and public actors that develop or deploy AI solutions to produce their own declarations for fear to be seen to be left behind, thus further contributing to the noise. And it risks creating a supermarket of principles and values, where private and public actors may shop for the kind of ethics that is best retrofitted to justify their behaviours, rather than revising their behaviours to make them consistent with a socially accepted ethical framework. However, the guidelines resolve these challenges because they are the closest thing available in the EU to a comprehensive and authoritative standard, offering a clear frame of reference and a common, conceptual vocabulary. They have been designed to establish a benchmark for what may or may not qualify, from now on, as trustworthy AI.

Further steps for a global stage

In some cases, a regulative approach may be premature, too prescriptive or stifle valuable innovation. An ethical approach leads to more flexible and still demanding expectations. It is important to remember that the publication of the guidelines is also just the first step. They will contribute to inform EU legislation and policies, but they also represent a roadmap for the rapid transformations enabled by AI technology[11]. In June 2019, the HLEG will issue its recommendations for the EU's AI research agenda, and on how the EU may strengthen its competitiveness in the development and deployment of AI, in line with the guidelines. And this summer, the EC will launch a pilot project to test the guidelines in collaboration with stakeholders to identify potential improvements and promote practical applications. The HLEG will review the outcome in early 2020 and further refine its output. In the long run, the EC "wants to bring this approach to AI ethics to the global stage … [and] strengthen cooperation with like-minded partners such as Japan, Canada or Singapore … [as well as] the G7 and G20"[12]. Some critics concede all this but still object that one cannot become a leader in ethical AI without becoming a leader in AI first[13,14]. Yet 'innovate first, fix later' is a mistake that, in the case of AI, could also be very costly and may cause a public backlash against AI, similar to the one against genetically modified crops in the past[15]. The climate change disaster and the trouble with social media platforms interfering in democracy should have taught us to plan innovation more carefully. This is why the EU wants to determine a long-term strategy in which ethics is an innovation enabler that offers a competitive advantage, and which ensures that fundamental rights and values are fostered, the public interest is served, and the natural environment thrives. Ethics-first is the right approach to set global standards for AI. The era of 'move fast and break things' is over. It is time to 'make haste slowly' (*festina lente*) in the development of AI. ❐

Luciano Floridi
University of Oxford, Oxford, UK.
e-mail: luciano.floridi@oii.ox.ac.uk

References

1. *Ethics Guidelines for Trustworthy AI* (European Commission, 2019); https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
2. *Draft Ethics Guidelines for Trustworthy AI* (European Commission, 2018); https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai
3. Floridi, L. et al. *Minds Mach.* **28**, 689–707 (2018).
4. Metzinger, T. *Der Tagesspiegel* https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html (2019).
5. Meyer, D. *Fortune* http://fortune.com/2019/04/08/eu-ai-ethics-principles/ (2019).
6. Floridi, L. *Philos. Trans. R. Soc. A* **376**, 20180081 (2018).
7. Wiener, N. *Science* **131**, 1355–1358 (1960).
8. Samuel, A. L. *Science* **132**, 741–742 (1960).
9. *AlgorithmWatch* https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/ (2019).
10. Winfield, A. *Alan Winfield's Web Log* http://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html (2019).
11. Floridi, L. & Lord Clement-Jones, T. *New Statesman* https://tech.newstatesman.com/policy/ai-ethics-framework (2019).
12. *Artificial Intelligence: Commission Takes Forward its Work on Ethics Guidelines* (European Commission, 2019); http://europa.eu/rapid/press-release_IP-19-1893_en.htm
13. Delcker, J. *Politico* https://www.politico.eu/article/europe-silver-bullet-global-ai-battle-ethics/ (2019).
14. Vincent, J. *The Verge* https://www.theverge.com/2019/4/8/18300149/eu-artificial-intelligence-ai-ethical-guidelines-recommendations (2019).
15. Cookson, C. *Financial Times* https://www.ft.com/content/0b301152-b0f8-11e8-99ca-68cf89602132 (2018).

Acknowledgements

L.F. is a member of the High-Level Expert Group (HLEG). Some of his research on the ethical impact of automation and algorithms has been funded by academic grants from the UK Research Council, the EU and Google Europe (also a member of the HLEG).