Harvard Data Science Review • Issue 1.1, Summer 2019

A Unified Framework of Five Principles for AI in Society

Luciano Floridi[1,2,3,4] **Josh Cowls**[4,5]

[1]**Digital Ethics Lab, Oxford Internet Institute, University of Oxford, Oxford, England, United Kingdom,**

[2]**Uehiro Centre for Practical Ethics, Faculty of Philosophy, Humanities Division, University of Oxford, England, United Kingdom,**

[3]**Department of Computer Science, University of Oxford, England, United Kingdom,**

[4]**Data Ethics Group, Alan Turing Institute, London, England, United Kingdom,**

[5]**Public Policy Programme, Alan Turing Institute, London, England, United Kingdom**

Published on: Sep 20, 2019

DOI: https://doi.org/10.1162/99608f92.8cd550d1

ABSTRACT

Artificial Intelligence (AI) is already having a major impact on society. As a result, many organizations have launched a wide range of initiatives to establish ethical principles for the adoption of socially beneficial AI. Unfortunately, the sheer volume of proposed principles threatens to overwhelm and confuse. How might this problem of 'principle proliferation' be solved? In this paper, we report the results of a fine-grained analysis of several of the highest-profile sets of ethical principles for AI. We assess whether these principles converge upon a set of agreed-upon principles, or diverge, with significant disagreement over what constitutes 'ethical AI.' Our analysis finds a high degree of overlap among the sets of principles we analyze. We then identify an overarching framework consisting of *five core principles* for ethical AI. Four of them are core principles commonly used in bioethics: *beneficence*, *non-maleficence*, *autonomy*, and *justice*. On the basis of our comparative analysis, we argue that a new principle is needed in addition: *explicability*, understood as incorporating both the epistemological sense of *intelligibility* (as an answer to the question 'how does it work?') and in the ethical sense of *accountability* (as an answer to the question: 'who is responsible for the way it works?'). In the ensuing discussion, we note the limitations and assess the implications of this ethical framework for future efforts to create laws, rules, technical standards, and best practices for ethical AI in a wide range of contexts.

Keywords: accountability, autonomy, artificial intelligence, beneficence, ethics, explicability, fairness, intelligibility, justice, non-maleficence.

---

1. Introduction

Artificial Intelligence (AI) is already having a major impact on society. The key questions are how, where, when, and by whom the impact of AI will be felt. As a result, many organizations have launched a wide range of initiatives to establish ethical principles for the adoption of socially beneficial AI. Unfortunately, the sheer volume of proposed principles threatens to become overwhelming and confusing, posing two potential problems.[1] Either the various sets of ethical principles for AI are similar, leading to unnecessary repetition and redundancy, or, if they differ significantly, confusion and ambiguity will result instead. The worst outcome would be a 'market for principles' where stakeholders may be tempted to 'shop' for the most appealing ones (Floridi, 2019b).

How might this problem of 'principle proliferation' be solved? In this paper, we report the results of a fine-grained analysis of several of the highest-profile sets of ethical principles for AI. We assess whether these principles are convergent, with a set of agreed-upon principles, or divergent, with significant disagreement over what constitutes 'ethical AI.' Our analysis finds a high degree of overlap among the sets of principles we analyze. We then identify an overarching framework consisting of *five core principles* for ethical AI. In the

ensuing discussion, we note the limitations and assess the implications of this ethical framework for future efforts to create laws, rules, standards, and best practices for ethical AI in a wide range of contexts.

2. Artificial Intelligence: A Research Area in Search of a Definition

AI has been defined in many ways. Today, it comprises several techno-scientific branches, well summarized in Figure 1 (see also the articles by Dick and Jordan in this issue for enlightening analyses).

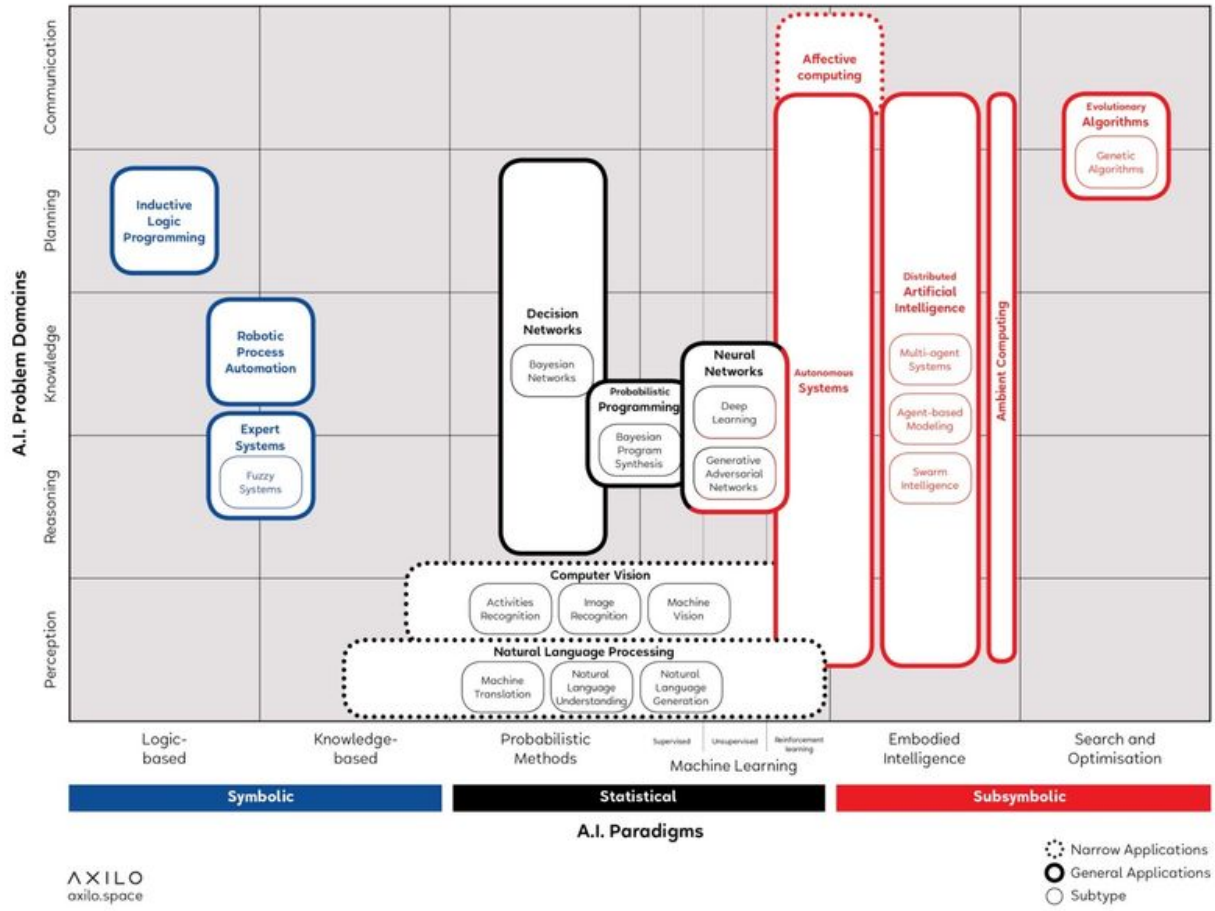

Figure 1. AI Knowledge Map (AIKM). Souce: Corea (2019). Reproduced with permission courtesy of F. Corea.

Altogether, AI paradigms still satisfy the classic definition provided by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon in their seminal *Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, the founding document and later event that established the new field of AI in 1955:

> For the present purpose the artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving. (Quotation from the 2006 re-issue in McCarthy et al., 2006 [1955]).

This is a *counterfactual*: *were* a human to behave in that way, that behaviour *would* be called intelligent. It does not mean that the machine *is* intelligent, or even *thinking*. The latter scenario is a fallacy, and smacks of superstition. Just because a dishwasher cleans the dishes as well as (or even better than) I do does not mean that it cleans them *like* I do, or needs any intelligence to achieve its task. The same counterfactual understanding of AI underpins the Turing test (Floridi, Taddeo, & Turilli, 2009), which, in this case, checks the ability of a machine to perform a task in such a way that the *outcome* would be indistinguishable from the outcome of a human agent working to achieve the same task (Turing, 1950).

The classic definition enables one to conceptualize AI as a growing resource of interactive, autonomous, and often self-learning agency (in the machine learning sense, see Figure 1), that can deal with tasks that would otherwise require human intelligence and intervention to be performed successfully. In short, AI is defined on the basis of engineered outcomes and actions and so, in what follows, we shall treat AI as a *reservoir of smart agency on tap* (see also Floridi 2019a). This is sufficiently general to capture the many ways in which AI is discussed in the documents we analyse in the rest of this article.

3. A Unified Framework of Five Principles for Ethical AI

The establishment of artificial intelligence as a field of academic research dates back to the 1950s (McCarthy et al., 2006 [1955]). The ethical debate is almost as old (Samuel, 1960; Wiener, 1960). However, it is only in recent years that impressive advances in the capabilities and applications of AI systems have brought the opportunities and risks of AI for society into sharper focus (Yang et al., 2018). The increasing demand for reflection and clear policies on the impact of AI on society has yielded a glut of initiatives. Each additional initiative yields a supplementary statement of principles, values, or tenets to guide the development and adoption of AI. The risk is unnecessary repetition and overlap, if the various sets of principles are similar, or confusion and ambiguity, if they differ. In either eventuality, the development of laws, rules, standards, and best practices to ensure that AI is socially beneficial may be delayed by the need to navigate the wealth of principles and declarations set out by an ever-expanding array of initiatives.

The time has come for a comparative analysis of these documents, including an assessment of whether they converge or diverge and, if the former, whether a unified framework may therefore be synthesised. For this comparative analysis, we identified six high-profile initiatives established in the interest of socially beneficial AI:

1. The Asilomar AI Principles, developed under the auspices of the Future of Life Institute, in collaboration with attendees of the high-level Asilomar conference of January 2017 (hereafter ‘Asilomar’; Asilomar AI Principles, 2017)
2. The Montreal Declaration for Responsible AI, developed under the auspices of the University of Montreal, following the Forum on the Socially Responsible Development of AI of November 2017 (hereafter ‘Montreal’; Montreal Declaration, 2017)[2]

3. The General Principles offered in the second version of *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. This crowd-sourced global treatise received contributions from 250 global thought leaders to develop principles and recommendations for the ethical development and design of autonomous and intelligent systems, and was published in December 2017 (hereafter 'IEEE'; IEEE, 2017, p. 6)[3]
4. The Ethical Principles offered in the *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*, published by the European Commission's European Group on Ethics in Science and New Technologies, in March 2018 (hereafter 'EGE'; EGE, 2018, pp. 16-20)
5. The 'five overarching principles for an AI code' offered in UK House of Lords Artificial Intelligence Committee's report, *AI in the UK: ready, willing and able?*, published in April 2018 (hereafter 'AIUK'; House of Lords, 2018, §417)
6. The Tenets of the Partnership on AI, a multi-stakeholder organization consisting of academics, researchers, civil society organisations, companies building and utilising AI technology, and other groups (hereafter 'the Partnership'; Partnership on AI, 2018).

Each set of principles meets three basic criteria: they are *recent*, published within the last three years; directly *relevant* to AI and its impact on society as a whole (thus excluding documents specific to a particular domain, industry, or sector); and highly *reputable*, published by authoritative, multi-stakeholder organizations with at least national scope.[4] Taken together, they yield 47 principles.[5] Overall, we find a degree of coherence and overlap between the six sets of principles that is impressive and reassuring. This convergence can most clearly be shown by comparing the sets of principles with the four core principles commonly used in bioethics: *beneficence*, *non-maleficence*, *autonomy*, and *justice* (Beauchamp & Childress, 2012). The comparison should not be surprising. Of all areas of applied ethics, bioethics is the one that most closely resembles digital ethics in dealing ecologically with new forms of agents, patients, and environments (Floridi, 2013). Yet while the four bioethical principles adapt surprisingly well to the fresh ethical challenges posed by artificial intelligence, they do not offer a perfect translation. As we shall see, the underlying meaning of each of the principles is contested, with similar terms often used to mean different things. Nor are the four principles exhaustive. On the basis of our comparative analysis, we argue that a new principle is needed in addition: *explicability*, understood as incorporating both *intelligibility* (for non-experts, e.g., patients or business customers, and for experts, e.g., product designers or engineers) and *accountability*. However, the convergence that we detect between these different sets of principles also demands caution. We explain the reasons for this caution in the following section, but first, we introduce the five principles.

3.1. Beneficence: Promoting Well-Being, Preserving Dignity, and Sustaining the Planet

The principle of creating AI technology that is beneficial to humanity is expressed in different ways across the six documents, but is perhaps the easiest of the four traditional bioethics principles to observe. Montreal and

IEEE principles both use the term "well-being"; for Montreal, "the development of AI should ultimately promote the well-being of all sentient creatures," while IEEE states the need to "prioritize human well-being as an outcome in all system designs." AIUK and Asilomar both characterise this principle as the "common good": AI should "be developed for the common good and the benefit of humanity," according to AIUK. The Partnership describes the intention to "ensure that AI technologies benefit and empower as many people as possible", while the EGE emphasizes the principle of both "human dignity" and "sustainability." Its principle of "sustainability" articulates perhaps the widest of all interpretations of beneficence, arguing that "AI technology must be in line with … ensur[ing] the basic preconditions for life on our planet, continued prospering for mankind and the preservation of a good environment for future generations." Taken together, the prominence of beneficence firmly underlines the central importance of promoting the well-being of people and the planet with AI.

3.2. Non-Maleficence: Privacy, Security and 'Capability Caution'

Though 'do only good' (beneficence) and 'do no harm' (non-maleficence) may seem logically equivalent, they are not, and represent distinct principles. While the six documents all encourage the creation of beneficent AI, each one also cautions against various negative consequences of overusing or misusing AI technologies (Cowls et al., 2018). Of particular concern is the prevention of infringements on personal privacy, which is included as a principle in five of the six sets. Several of the documents emphasize avoiding the misuse of AI technologies in other ways. The Asilomar Principles warn against the threats of an AI arms race and of the recursive self-improvement of AI, while the Partnership similarly asserts the importance of AI operating "within secure constraints." The IEEE document meanwhile cites the need to "avoid misuse," and the Montreal Declaration argues that those developing AI "should assume their responsibility by working against the risks arising from their technological innovations." Yet from these various warnings, it is not entirely clear whether it is the people developing AI, or the technology itself, which should be encouraged not to do harm; in other words, whether it is Frankenstein or his monster against whose maleficence we should be guarding. At the heart of this quandary is the question of autonomy.

3.3. Autonomy: The Power to Decide (to Decide)

When we adopt AI and its smart agency, we willingly cede some of our decision-making power to technological artefacts. Thus, affirming the principle of autonomy in the context of AI means striking a balance between the decision-making power we retain for ourselves and that which we delegate to artificial agents. The risk is that the growth in *artificial autonomy* may undermine the flourishing of *human autonomy*. It is not therefore surprising that the principle of autonomy is explicitly stated in four of the six documents. The Montreal Declaration articulates the need for a balance between human- and machine-led decision-making, stating that "the development of AI should promote the *autonomy* [italics added] of all human beings". The EGE argues that autonomous systems "must not impair [the] freedom of human beings to set their own standards and norms," while AIUK adopts the narrower stance that "the autonomous power to hurt, destroy or

deceive human beings should never be vested in AI." The Asilomar document similarly supports the principle of autonomy, insofar as "humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives." It is therefore clear both that the autonomy of humans should be promoted and that the autonomy of machines should be restricted and made intrinsically reversible, should human autonomy need to be protected or re-established (consider the case of a pilot able to turn off the automatic pilot and regain full control of the airplane). This introduces a notion we might call 'meta-autonomy,' or a 'decide-to-delegate' model: humans should retain the power to decide which decisions to take: exercising the freedom to choose where necessary, and ceding it in cases where overriding reasons, such as efficacy, may outweigh the loss of control over decision-making. Any delegation should also remain overridable in principle (i.e., deciding to decide again).

3.4. Justice: Promoting Prosperity, Preserving Solidarity, Avoiding Unfairness

The decision to make or delegate decisions does not take place in a vacuum. Nor is this capacity distributed equally across society. The consequences of this disparity in autonomy are addressed in the principle of justice. The importance of 'justice' is explicitly cited in the Montreal Declaration, which argues that "the development of AI should promote justice and seek to eliminate all types of discrimination," while the Asilomar Principles include the need for both "shared benefit" and "shared prosperity" from AI. Under its principle named "Justice, equity and solidarity," the EGE argues that AI should "contribute to global justice and equal access to the benefits" of AI technologies. It also warns against the risk of bias in datasets used to train AI systems, and – unique among the documents – argues for the need to defend against threats to "solidarity," including "systems of mutual assistance such as in social insurance and healthcare." Elsewhere 'justice' has still other meanings (especially in the sense of *fairness*), variously relating to the use of AI to correct past wrongs such as eliminating unfair discrimination, promoting diversity, and preventing the rise of new threats to justice. The diverse ways in which justice is characterised hints at a broader lack of clarity over AI as a human-made reservoir of 'smart agency.' Put simply, are we (humans) the patient, receiving the 'treatment' of AI, the doctor prescribing it? Or both? This question can only be resolved with the introduction of a fifth principle which emerges from our analysis.

3.5. Explicability: Enabling the Other Principles through Intelligibility and Accountability

The short answer to the question of whether 'we' are the patient or the doctor is that actually we could be either, depending on the circumstances and on who 'we' are in everyday life. The situation is inherently unequal: a small fraction of humanity is currently engaged in the development of a set of technologies that are already transforming the everyday lives of almost everyone else. This stark reality is not lost on the authors whose documents we analyze. All of them refer to the need to understand and hold to account the decision-making processes of AI. Different terms express this principle: "transparency" in Asilomar and EGE; both "transparency" and "accountability" in IEEE; "intelligibility" in AIUK; and as "understandable and

interpretable" by the Partnership. Each of these principles captures something seemingly novel about AI: that its workings are often invisible or unintelligible to all but (at best) the most expert observers.

The addition of the principle of 'explicability,' incorporating both the epistemological sense of 'intelligibility' (as an answer to the question 'how does it work?') and in the ethical sense of 'accountability' (as an answer to the question 'who is responsible for the way it works?'), is the crucial missing piece of the AI ethics jigsaw. It complements the other four principles: for AI to be beneficent and non-maleficent, we must be able to understand the good or harm it is actually doing to society, and in which ways; for AI to promote and not constrain human autonomy, our 'decision about who should decide' must be informed by knowledge of how AI would act instead of us; and for AI to be just, we must know whom to hold accountable in the event of a serious, negative outcome, which would require in turn adequate understanding of why this outcome arose.

3.6. A Synoptic View

Taken together, these five principles capture every one of the 47 principles contained in the six high-profile, expert-driven documents we analysed. Moreover, each principle is included in almost every statement of principles we analyzed (see Table 1 below). The five principles therefore form an ethical framework within which policies, best practices, and other recommendations may be made. This framework of principles is shown in Figure 2.

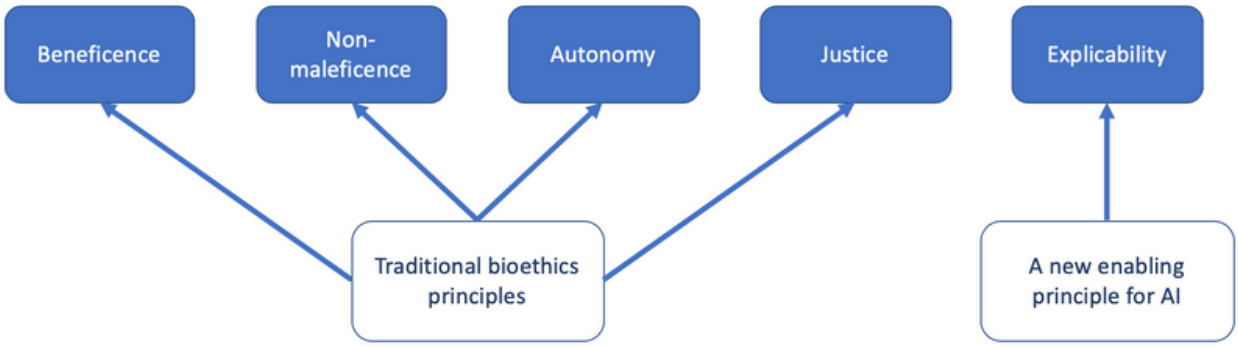

Figure 2: An ethical framework of the five overarching principles for AI which emerged from the analysis.

4. AI Ethics: Whence and For Whom?

It is important to note that each of the six sets of ethical principles for AI that we analyzed emerged either from initiatives with global scope, or from within western liberal democracies. For the framework to be more broadly applicable, it would undoubtedly benefit from the perspectives of regions and cultures presently un- or under-represented in our sample. Of particular interest in this respect is the role of China, which is already home to the world's most valuable AI start-up (Jezard, 2018), enjoys various structural advantages in developing AI (Lee & Triolo, 2017), and whose government has stated its ambitions to lead the world in state-

of-the-art AI technology by 2030 (China State Council, 2017). In its State Council Notice on AI and elsewhere, the Chinese government has expressed interest in further consideration of the social and ethical impact of AI (Ding, 2018; Webster et. al, 2017). Nor is enthusiasm about the use of technologies unique to governments, but it is also shared by general publics – more so those in China and India than in Europe or the USA, as new representative survey research shows (Vodafone Institute, 2018).

An executive at the major Chinese technology firm Tencent recently suggested that the European Union should focus on developing AI which has "the maximum benefit for human life, even if that technology isn't competitive to take on [the] American or Chinese market" (Boland, 2018). This has been echoed by claims that ethics may be "Europe's silver bullet" in the "global AI battle" (Delcker, 2018). We disagree. Ethics is not the preserve of a single continent or culture. Every company, government agency, and academic institution designing, developing or deploying AI has an obligation to do so in line with an ethical framework along the lines of the one we present here, broadened to incorporate a more geographically, culturally, and socially diverse array of perspectives (Cowls et al., n.d.). Similarly, laws, rules, standards and best practices to constrain or control AI – including all those currently under consideration by regulatory bodies, legislatures and industry groups –would also benefit from close engagement with a unified framework of ethical principles.

5. Conclusion: From Principles to Practices

If the framework presented in this article provides a coherent and sufficiently comprehensive overview of the central ethical principles for AI (Floridi et al., 2018), then it can serve as the architecture within which laws, rules, technical standards, and best practices are developed for specific sectors, industries, and jurisdictions. In these contexts, the framework may play both an enabling role (consider, for example, the use of AI to help meet the United Nations Sustainable Development Goals), and a constraining one (as in the need to regulate AI technologies in the context of online crime and cyberwarfare: King et al., 2018; Taddeo & Floridi, 2018b). Indeed, the framework played a valuable role in the work of AI4People, Europe's first global forum on the social impact of AI, which recently adopted it to propose 20 concrete recommendations for a 'Good AI Society' to the European Commission (Floridi et al., 2018). Since then it has been largely adopted by the *Ethics Guidelines for Trustworthy AI* published by the European Commission's High-Level Expert Group on Artificial Intelligence (HLEGAI 2018 and 2019), which in turn has influenced the OECD's *Recommendation of the Council on Artificial Intelligence* (OECD 2019), reaching 42 countries[6] (see Table 1).

Table 1: The five principles in the six documents analyzed and their occurrence in three recent documents

	Beneficence	*Nonmaleficence*	*Autonomy*	*Justice*	*Explicability*
AIUK	•	•	•	•	•

Asilomar	•	•	•	•	•
EGE	•	•	•	•	•
IEEE	•	•			•
Montreal	•	•	•	•	•
Partnership	•	•		•	•
AI4People	•	•	•	•	•
EC HLEG	•	•	•	•	•
OECD	•	•	•	•	•

The development and use of AI hold the potential for both positive and negative impact on society, to alleviate or to amplify existing inequalities, to cure old problems, or to cause new ones. Charting the course that is socially preferable will depend not only on well-crafted regulation and common standards, but also on the use of a framework of ethical principles, within which concrete actions can be situated. We believe that the framework presented here as emerging from the current debate will serve as valuable architecture for securing positive social outcomes from AI technology and move from good principles to good practices (Cowls et al., 2019; Morley et al., 2019).

Acknowledgements

Floridi's work was supported by (i) Privacy and Trust Stream - Social lead of the PETRAS Internet of Things research hub - PETRAS is funded by the UK Engineering and Physical Sciences Research Council (EPSRC), grant agreement no. EP/N023013/1; (ii) Facebook; and (iii) Google. Cowls is the recipient of a Doctoral Studentship from the Alan Turing Institute.

Disclosure Statement

Floridi chaired the AI4People project and Cowls was the rapporteur. Floridi is also a member of the European Commission's High-Level Expert Group on Artificial Intelligence (HLEGAI).

References

Beauchamp, T. L., & Childress, J. F. (2012). *Principles of biomedical ethics*. Oxford: Oxford University Press.

Boland, H. (2018, October, 14). Tencent executive urges Europe to focus on ethical uses of artificial intelligence. *The Telegraph*. https://www.telegraph.co.uk/technology/2018/10/14/tencent-executive-urges-europe-focus-ethical-uses-artificial/

China State Council (2017, July, 8). *State council notice on the issuance of the next generation artificial intelligence development plan*. Retrieved September 18, 2018, from http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm. Trans by Creemers, R., Webster, G., Triolo, P. and Kania, E. https://www.newamerica.org/documents/1959/translation-fulltext-8.1.17.pdf

Corea, F. (2019). AI knowledge map: How to classify AI technologies, a sketch of a new AI technology landscape. First appeared in *Medium.* https://medium.com/@Francesco_AI/ai-knowledge-map-how-to-classify-ai-technologies-6c073b969020 reproduced in Corea, F. (2019). In *An Introduction to Data* (p. 26). Springer.

Cowls, J., Floridi, L., & Taddeo, M. (2018). The challenges and opportunities of ethical AI. *Artificially Intelligent.* https://digitransglasgow.github.io/ArtificiallyIntelligent/contributions/04_Alan_Turing_Institute.html

Cowls, J., King, T. C., Taddeo, M., & Floridi, L. (2019). Designing AI for social good: Seven essential factors. *SSRN*. https://doi.org/10.2139/ssrn.3388669

Cowls, J., Png, M.-T., & Au, Y. (n.d.). Foundations for geographic representation in algorithmic ethics. Unpublished.

Delcker, J. (2018, March, 3). Europe's silver bullet in global AI battle: Ethics. *Politico*. https://www.politico.eu/article/europe-silver-bullet-global-ai-battle-ethics/

Ding, J. (2018, March). Deciphering China's AI dream. https://www.fhi.ox.ac.uk/wp-content/uploads/Deciphering_Chinas_AI-Dream.pdf

European Group on Ethics in Science and New Technologies (2018, March). *Statement on artificial intelligence, robotics and 'autonomous' systems.* https://ec.europa.eu/info/news/ethics-artificial-intelligence-statement-ege-released-2018-apr-24_en

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds & Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Taddeo, M., & Turilli, M. (2009). Turing's imitation game: still an impossible challenge for all machines and some judges—an evaluation of the 2008 Loebner contest. *Minds and Machines*, *19*(1), 145–150. https://doi.org/10.1007/s11023-008-9130-6

Floridi, L. (2013). *The ethics of information*. Oxford, Oxford University Press.

Floridi, L. (2019a). What the near future of artificial intelligence could be. *Philosophy & Technology, 32*(1), 1–15. https://doi.org/10.1007/s13347-019-00345-y

Floridi, L. (2019b). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, *32*(2), 185–193. https://doi.org/10.1007/s13347-019-00354-x

Hagendorff, T. (2019). The ethics of AI ethics—An evaluation of guidelines. *arXiv*. https://doi.org/10.48550/arXiv.1903.03425

HLEGAI [High Level Expert Group on Artificial Intelligence], European Commission (2018, December 18). *Draft ethics guidelines for trustworthy AI*. https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai

HLEGAI [High Level Expert Group on Artificial Intelligence], European Commission (2019, April 8). *Ethics guidelines for trustworthy AI* https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

House of Lords Artificial Intelligence Committee (2018, April 16). *AI in the UK: ready, willing and able*. https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm

The IEEE Initiative on Ethics of Autonomous and Intelligent Systems (2017). Ethically Aligned Design, v2. https://ethicsinaction.ieee.org

Jezard, A. (2018, April 11). *China is now home to the world's most valuable AI start-up*. World Economic Forum. https://www.weforum.org/agenda/2018/04/chart-of-the-day-china-now-has-the-worlds-most-valuable-ai-startup/

King, T., Aggarwal, N., Taddeo, M., & Floridi, L. (2018, May 22). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *SSRN*. https://doi.org/10.2139/ssrn.3183238

Lee, K., & Triolo, P. (2017, December). China's artificial intelligence revolution: Understanding Beijing's structural advantages. *Eurasian Group*. https://www.eurasiagroup.net/live-post/ai-in-china-cutting-through-the-hype

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12. https://doi.org/10.1609/aimag.v27i4.1904

Montreal Declaration for a Responsible Development of Artificial Intelligence (2017, November, 3). Announced at the conclusion of the Forum on the Socially Responsible Development of AI.

https://www.montrealdeclaration-responsibleai.com/the-declaration

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*. Retrieved from https://doi.org/10.48550/arXiv.1905.06876

OECD (2019). Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Partnership on AI (2018). Tenets. https://www.partnershiponai.org/tenets/

Samuel, A. L. (1960). Some moral and technical consequences of automation—A refutation. *Science, 132*(3429), 741–742. https://doi.org/10.1126/science.132.3429.741

Taddeo, M., & Floridi, L. (2018a). How AI can be a force for good. *Science*, *361*(6404), 751–752. http://doi.org/10.1126/science.aat5991

Taddeo, M., & Floridi, L. (2018b). Regulate artificial intelligence to avert cyber arms race. *Nature*, *556*(7701), 296–298. https://doi.org/10.1038/d41586-018-04602-6

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 5*(236), 433–460. https://doi.org/10.1093/mind/lix.236.433

Vodafone Institute for Society and Communications (2018). *New technologies: India and China see enormous potential—Europeans more sceptical*. https://www.vodafone-institut.de/digitising-europe/digitisation-india-and-china-see-enormous-potential/

Webster, G., Creemers, R., Triolo, P. and Kania, E (2017, August 1). China's plan to 'Lead' in AI: Purpose, prospects, and problems. *New America*. https://www.newamerica.org/cybersecurity-initiative/blog/chinas-plan-lead-ai-purpose-prospects-and-problems/

Wiener, N. (1960). Some moral and technical consequences of automation. *Science, 131*(3410), 1355–1358. https://doi.org/10.1126/science.131.3410.1355

Yang, G. Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., Jacobstein, N., Kumar, V., McNutt, M., Merrifield, R., Nelson, B. J., Scassellati, B., Taddeo, M., Taylor, R., Veloso, M., Wang, Z. L., & Wood, R. (2018). The grand challenges of Science Robotics. *Science Robotics*, *3*(14), Article eaar7650. https://doi.org/10.1126/scirobotics.aar7650

---

the article.

Footnotes

1. These are not the only problems, see (Floridi 2019b). ↩

2. The Montreal Declaration is currently open for comments as part of a redrafting exercise. The principles we refer to here are those which were publicly announced as of May 1, 2018. ↩

3. The third version of *Ethically Aligned Design* will be released in 2019 following wider public consultation. ↩

4. A similar evaluation of AI ethics guidelines has recently been undertaken by Hagendorff (2019), which adopts different criteria of inclusion and assessment. Note that the evaluation includes in its sample the set of principles we describe here. ↩

5. Of the six documents, the Asilomar Principles offer the largest number of principles with arguably the broadest scope. The 23 principles are organised under three headings, “research issues”, “ethics and values”, and “longer-term issues”. We have omitted consideration of the five “research issues” here as they are related specifically to the practicalities of AI development in the narrower context of academia and industry. Similarly, the Partnership’s eight Tenets consist of both intra-organisational objectives and wider principles for the development and use of AI. We include only the wider principles (the first, sixth, and seventh tenets). ↩

6. https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm ↩