

A Unified Framework of Ethical Principles for AI

4.0 Summary

Previously, in Chapters 1–3, we saw how AI is a new form of agency that can deal with tasks and problems successfully, in view of a goal, without any need for intelligence. Every success of any AI application does not move the bar of what it means to be an intelligent agent. Instead, it bypasses the bar altogether. The success of such artificial agency is increasingly facilitated by the *enveloping* (that is, reshaping into AI-friendly contexts) of the environments in which AI operates. The decoupling of agency and intelligence and the enveloping of the world generate significant ethical challenges, especially in relation to autonomy, bias, explainability, fairness, privacy, responsibility, transparency, and trust (yes, mere alphabetic order). For this reason, many organizations launched a wide range of initiatives to establish ethical principles for the adoption of socially beneficial AI after the Asilomar AI Principles and the Montreal Declaration for a Responsible Development of Artificial Intelligence were published in 2017. This soon became a cottage industry. Unfortunately, the sheer volume of proposed principles threatens to overwhelm and confuse.

This chapter presents the results of a comparative analysis of several of the highest profile sets of ethical principles for AI. I assess whether these principles converge around a set of agreed-upon principles, or diverge with significant disagreement over what constitutes ‘ethical AI’. In the chapter, I argue that there is a high degree of overlap across the sets of principles under analysis. There is an overarching framework consisting of *five core principles* for ethical AI. Four of them are core principles commonly used in bioethics: *beneficence*, *nonmaleficence*, *autonomy*, and *justice*. This is unsurprising when AI is interpreted as a form of agency. And the coherence between bioethics and the ethics of AI may be seen as evidence that approaching AI as a new form of agency is fruitful. Readers who disagree may find that an intelligent-oriented interpretation of AI could cohere more easily with a virtue ethics approach, rather than with bioethics. I find approaching the ethics of AI and digital ethics in general from a virtue ethics perspective unconvincing, but it remains a viable possibility and this is not the place to rehearse my objections (Floridi 2013).

Based on the comparative analysis, the addition of a new principle is needed: *explicability*. Explicability is 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?’ The extra principle is required by the fact that AI is a *new* form of agency. The chapter ends with a discussion of 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.

4.1 Introduction: Too Many Principles?

As we have seen in the preceding chapters, AI is already having a major impact on society, which will only increase in the future. The key questions are *how*, *where*, *when*, and *by whom* the impact of AI will be felt. I shall return to this point in Chapter 10. Here I would like to focus on a consequence of this trend: 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—already more than 160 in 2020, according to Algorithm Watch’s AI Ethics Guidelines Global Inventory (Algorithm Watch 2020)¹—threatens to overwhelm and confuse. This poses two potential problems: if the various sets of ethical principles for AI are similar, this leads to unnecessary repetition and redundancy; yet if they differ significantly, confusion and ambiguity will result instead. The worst outcome would be a ‘market for principles’ wherein stakeholders may be tempted to ‘shop’ for the most appealing ones, as we shall see in Chapter 5.

How might this problem of ‘principle proliferation’ be solved? In this chapter, I present and discuss the results of a comparative analysis of several of the highest profile sets of ethical principles for AI. I assess whether these principles converge around a shared set of agreed-upon principles, or diverge with significant disagreement over what constitutes ethical AI. The analysis finds a high degree of overlap among the sets of principles under analysis. This leads to the identification of an overarching framework consisting of *five core principles* for ethical AI. In the ensuing discussion, I 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.

4.2 A Unified Framework of Five Principles for Ethical AI

We saw in Chapter 1 that the establishment of AI as a field of academic research dates to the 1950s (McCarthy et al. 2006 [1955]). The ethical debate is almost as old (Wiener 1960, Samuel 1960b). But 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

¹ See also Jobin, Ienca, and Vayena (2019).

increasing demand for reflection and clear policies on the impact of AI on society has yielded a glut of initiatives. Each additional initiative provides a supplementary statement of principles, values, or tenets to guide the development and adoption of AI. You get the impression that a ‘me too’ escalation had taken place at some point. No organization could be without some ethical principles for AI and accepting the ones already available looked inelegant, unoriginal, or lacking in leadership.

The ‘me too’ attitude was followed up by ‘mine and only mine.’ Years later, the risk is still 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 synthesized. Table 4.1 shows the six high-profile initiatives established in the interest of socially beneficial AI chosen as the most informative for a comparative analysis. I shall analyse two more sets of principles later in the chapter.

Each set of principles in Table 4.1 meets four basic criteria. They are:

Table 4.1 Six sets of ethical principles for AI

1) The Asilomar AI Principles (Future of Life Institute 2017) developed under the auspices of the Future of Life Institute in collaboration with attendees of the high-level Asilomar conference in January 2017 (hereafter Asilomar);
2) The Montreal Declaration (Université de Montréal 2017) developed under the auspices of the University of Montreal following the Forum on the Socially Responsible Development of AI in November 2017 (hereafter Montreal); 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 (IEEE (2017), 6). This crowd-sourced document 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);
4) The ethical principles offered in the Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems (EGE 2018) published by the European Commission’s European Group on Ethics in Science and New Technologies in March 2018 (hereafter EGE);
5) The ‘five overarching principles for an AI code’ offered in the UK House of Lords Artificial Intelligence Select Committee’s report ‘AI in the UK: Ready, Willing and Able?’ (House of Lords Artificial Intelligence Committee 16 April 2017, §417) published in April 2018 (hereafter AIUK);
6) The Tenets of the Partnership on AI (Partnership on AI 2018) published by a multi-stakeholder organization consisting of academics, researchers, civil society organizations, companies building and utilizing AI technology, and other groups (hereafter the Partnership).

² The principles referred to here are those that were publicly announced as of 1 May 2018 and available at the time of writing this chapter here: <https://www.montrealdeclaration-responsibleai.com/the-declaration>.

- a) *recent*, as in published since 2017;
- b) directly *relevant* to AI and its impact on society as a whole (thus excluding documents specific to a particular domain, industry, or sector);
- c) highly *reputable*, published by authoritative, multi-stakeholder organizations with at least a national scope;³ and
- d) *influential* (because of a–c).

Taken together, they yield forty-seven principles.⁴ Despite this, the differences are mainly linguistic and there is a degree of coherence and overlap across the six sets of principles that is impressive and reassuring. The convergence can most clearly be shown by comparing the sets of principles with the four core principles commonly used in bioethics: *beneficence*, *nonmaleficence*, *autonomy*, and *justice* (Beauchamp and Childress 2013). This comparison should not be surprising. As the reader may recall, the view I support in this book is that AI is not a new form of intelligence but an unprecedented form of agency. So of all the areas of applied ethics, bioethics is the one that most closely resembles digital ethics when dealing ecologically with new forms of agents, patients, and environments (Floridi 2013).

The four bioethical principles adapt surprisingly well to the fresh ethical challenges posed by AI. However, 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. Based on the comparative analysis, it becomes clear, as noted in the summary, that an additional principle is needed: *explicability*. *Explicability* is 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?’ This additional principle is needed both for experts, such as product designers or engineers, and for non-experts, such as patients or business customers.⁵ And the new principle is required by the fact that AI is not like any biological form of agency. However, the convergence detected between these different sets of principles also demands caution. I explain the reasons for this caution later in the chapter, but first, let me introduce the five principles.

³ A similar evaluation of AI ethics guidelines, undertaken recently by Hagendorff (2020), adopts different criteria for inclusion and assessment. Note that the evaluation includes in its sample the set of principles we describe here.

⁴ Of the six documents, the Asilomar Principles offer the largest number of principles with arguably the broadest scope. The twenty-three principles are organized under three headings: ‘research issues’, ‘ethics and values’, and ‘longer-term issues’. Consideration of the five ‘research issues’ is omitted 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-organizational objectives and wider principles for the development and use of AI. Only the wider principles (the first, sixth, and seventh tenets) are included here.

⁵ There is a third question that concerns what new insights can be provided by explicability. It is not relevant in this context, but equally important; for detail, see Watson and Floridi (2020).

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

The principle of creating AI technologies that are beneficial to humanity is expressed in different ways across the six documents. Still, it is perhaps the easiest of the four traditional bioethics principles to observe. The Montreal and IEEE principles both use the term ‘well-being’; Montreal notes that ‘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 characterize 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. These are points to which I shall return in the following chapters.

4.4 Nonmaleficence: Privacy, Security, and ‘Capability Caution’

Although ‘do only good’ (beneficence) and ‘do no harm’ (nonmaleficence) may seem logically equivalent, they are not. Each represents a distinct principle. The six documents all encourage the creation of beneficent AI, and each one also cautions against the various negative consequences of overusing or misusing AI technologies (see Chapters 5, 7, and 8). 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, that should be encouraged to do no harm. In other words, it is unclear whether it is Dr Frankenstein (as I suggest) or his monster against whose maleficence we should be guarding. At the heart of this quandary is the question of autonomy.

4.5 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 AAs. The risk is that the growth in *artificial autonomy* may undermine the flourishing of *human autonomy*. It is therefore unsurprising 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* [emphasis 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. The latter should be 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 function and regain full control of the airplane). This introduces a notion one may call *meta-autonomy*, or a *decide-to-delegate* model. Humans should retain the power to decide which decisions to take and in what priority, 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. But any delegation should also remain overridable in principle, adopting an ultimate safeguard of *deciding to decide again*.

4.6 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 by 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’. Likewise, 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 related to the use of AI in correcting past wrongs, such as by eliminating unfair discrimination, promoting diversity, and preventing the reinforcement of biases or the rise of new threats to justice. The diverse ways in which justice is characterized 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, which would be acting as 'the doctor' prescribing it, or both? This question can only be resolved with the introduction of a fifth principle that emerges from the previous analysis.

4.7 Explicability: Enabling the Other Principles through Intelligibility and Accountability

In response to the previous question about whether we are the patient or the doctor, the short answer is that we could actually be either. It depends on the circumstances and on who 'we' refers to 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 of the documents under analysis. 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 as a form of agency: that its workings are often invisible, opaque, or unintelligible to all but (at best) the most expert observers.

The principle of 'explicability' incorporates both the epistemological sense of 'intelligibility' and the ethical sense of 'accountability'. The addition of this principle is the crucial missing piece of the AI ethics jigsaw. It complements the other four principles: in order for AI to be beneficent and non-maleficent, we must be able to understand the good or harm that it is actually doing to society, and in what 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 how to improve its performance; and for AI to be just, we must know whom to hold ethically responsible (or actually legally liable) in the event of a serious, negative outcome, which would in turn require an adequate understanding of why this outcome arose, and how it could be prevented or minimized in the future.

4.8 A Synoptic View

Taken together, the previous five principles capture every one of the forty-seven principles contained in the six high-profile, expert-driven documents analysed in Table 4.1. As a test, it is worth stressing that each principle is included in almost every statement of principles analysed in Table 4.2 (see Section 4.10). 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 4.1.

4.9 AI Ethics: Whence and for Whom?

It is important to note that each of the six sets of ethical principles for AI in Table 4.1 emerged either from initiatives with a 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 11 April 2018), enjoys various structural advantages in developing AI (Lee and Triolo December 2017), and whose government has stated its ambitions to lead the world in state-of-the-art AI technology by 2030 (China State Council 8 July 2017). This is not the context for an analysis of China’s AI policies,⁶ so let me add only a few closing remarks.

In its State Council Notice on AI and elsewhere, the Chinese government expressed interest in further consideration of the social and ethical impact of AI (Ding 2018). Enthusiasm for the use of technologies is hardly unique to governments. It is also shared by the general public—more so those in China and India than in Europe or the USA, as new representative survey research shows

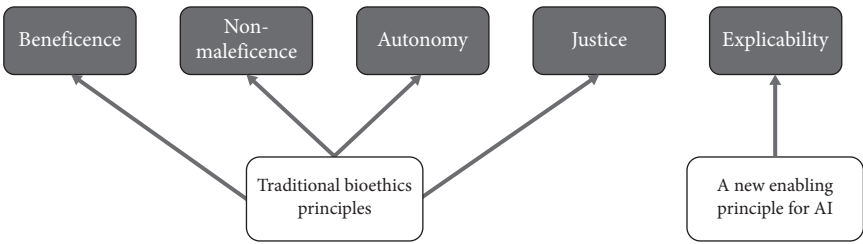


Figure 4.1 An ethical framework of the five overarching principles for AI

⁶ On this, see Roberts, Cowls, Morley, et al. (2021), Roberts, Cowls, Hine, Morley, et al. (2021), Hine and Floridi (2022).

(Vodafone Institute for Society and Communications 2018). In the past, an executive at the major Chinese technology firm Tencent suggested that the EU should focus on developing AI that has ‘the *maximum* benefit for human life, even if that technology isn’t competitive to take on [the] American or Chinese market’ (Boland 14 October 2018). This was echoed by claims that ethics may be ‘Europe’s silver bullet’ in the ‘global AI battle’ (Delcker 3 March 2018). I 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—even if not along the lines of the one presented here—broadened to incorporate a more geographically, culturally, and socially diverse array of perspectives. 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) could also benefit from close engagement with a unified framework of ethical principles. What remains true is that the EU may have an advantage in the so-called Brussels effect, but this only places more responsibility on how AI is regulated by the EU.

4.10 Conclusion: From Principles to Practices

If the framework presented in this chapter 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 (see Chapter 12) and a constraining one. The latter role refers to the need to regulate AI technologies in the context of online crime (see Chapter 7) and cyberwar, which I discuss in Floridi and Taddeo (2014), Taddeo and Floridi (2018b). Indeed, the framework shown in Figure 4.1 played a valuable role in five other documents:

- 1) the work of AI4People (Floridi et al. 2018), Europe’s first global forum on the social impact of AI, which adopted it to propose twenty concrete recommendations for a ‘Good AI Society’ to the European Commission (disclosure: I chaired the project; see ‘AI4People’ in Table 4.2).
The work of AI4People was largely adopted by
- 2) the Ethics Guidelines for Trustworthy AI published by the European Commission’s High-Level Expert Group on AI (HLEGAI 18 December 2018, 8 April 2019) (disclosure: I was a member; see ‘HLEG’ in Table 4.2); which in turn influenced
- 3) the OECD’s Recommendation of the Council on Artificial Intelligence (OECD 2019); see ‘OECD’ in Table 4.2), which reached forty-two countries; and

Table 4.2 The five principles in the analysed documents

	Beneficence	Nonmaleficence	Autonomy	Justice	Explicability
AIUK	•	•	•	•	•
Asilomar	•	•	•	•	•
EGE	•	•	•	•	•
IEEE	•	•			•
Montreal	•	•	•	•	•
Partnership	•	•		•	•
AI4People	•	•	•	•	•
HLEG	•	•	•	•	•
OECD	•	•	•	•	•
Beijing	•	•		•	•
Rome Call	•	•	•	•	•

4) The EU AI Act.

All this was followed by the publication of the so-called

- 5) Beijing AI Principles (Beijing Academy of Artificial Intelligence 2019); see ‘Beijing’ in Table 4.2) and
- 6) the so-called Rome Call for an AI Ethics (Pontifical Academy for Life 2020), a document elaborated by the Pontifical Academy for Life, which co-signed it with Microsoft, IBM, FAO, and the Italian Ministry of Innovation (disclosure: I was among the authors of the draft; see ‘Rome Call’ in Table 4.2).

The development and use of AI hold the potential for both positive and negative impacts on society, to alleviate or to amplify existing inequalities, to solve old and new problems, or to cause unprecedented ones. Charting the course that is socially preferable (equitable) and environmentally sustainable 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. The framework presented in this chapter as emerging from current debate may serve as a valuable architecture for securing positive social outcomes from AI technology. It can help move from good principles to good practices (Floridi et al. 2020, Morley, Floridi, et al. 2020). It can also provide the framework required by an ethics-based auditing of AI (Mokander and Floridi 2021, Floridi et al. 2022, Mökander et al. forthcoming). During this translation, some risks need to be mapped to ensure that the ethics of AI does not fall into new or old traps. This is the task of the next chapter.