Prolegomena to a White Paper on an Ethical Framework for a Good AI Society

Josh Cowls and Luciano Floridi

1. Introduction

That AI will have a major impact on society is no longer in question. Current debate turns instead on how far this impact will be positive or negative, for whom, in which ways, in which places, and on what timescale. Put another way, we can safely dispense with the question of *whether* AI will have an impact; the pertinent questions now are *by whom, how*, *where*, and *when* this positive or negative impact will be felt, and hence what governance needs to be put in place to provide the best possible answers.

In order to frame these questions in a more substantive way, in this prolegomena we introduce what we consider the four core opportunities for society offered by the *use* of AI, four associated risks which could emerge from its *overuse* or *misuse*, and the opportunity costs associated with its *underuse*. We then offer a high-level view of the emerging advantages for organisations of taking an ethical approach to developing and deploying AI. Finally, we introduce a set of five principles which should guide the development and deployment of AI technologies – four of which build on existing bioethics principles and an additional one that we argue is of equal importance in the case of AI.

2. The Opportunities and Risks of AI for Society

In this section, we introduce what we consider the four chief opportunities for society that AI offers. They are four because they address the four fundamental points in the understanding of human dignity and flourishing: *who we can become* (autonomous self-realisation); *what we can do* (human agency); *what we can achieve* (societal capabilities); and *how we can interact with each other and the world* (societal cohesion). In each case, AI can be *used* to foster human nature and its potentialities, thus creating opportunities, *underused*, thus creating opportunity costs, or *overused* and *misused*, thus creating risks. As the terminology indicates, the assumption is that the *use* of AI is synonymous with good innovation and positive applications of this technology. However, fear, ignorance, misplaced concerns or excessive reaction may lead a society to *underuse* AI technologies below their full potential, for what might be broadly described as the wrong reasons. This might include, for

example, heavy-handed or misconceived regulation, under-investment, or a public backlash akin to that faced by genetically modified crops (Imperial College, 2017). As a result, the benefits offered by AI technologies may not be fully realised by society. These dangers arise largely from unintended consequences and relate typically to good intentions gone awry. However, we must also consider the risks associated with inadvertent *overuse* or wilful *misuse* of AI technologies, grounded, for example, in misaligned incentives, adversarial geopolitics, greed, or malicious intent. Everything from email scams to full-scale cyber-warfare may be accelerated or intensified by the malicious use of AI technologies (Taddeo, 2017). And new evils may be made possible (King et. al, 2018). The possibility of social progress represented by the aforementioned opportunities must be weighed against the risk that malicious manipulation will be enabled or enhanced by AI. Yet a broad risk is that AI may be underused out of fear of overuse or misuse. We summarise these risks in Figure A below, and offer a more detailed explanation in the text that follows.

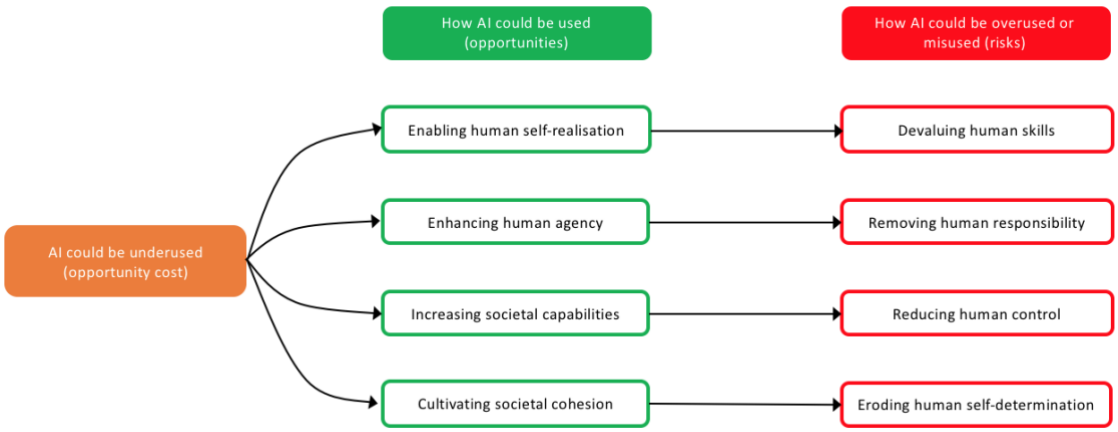

Figure A: Overview of the four core opportunities offered by AI, four corresponding risks, and the opportunity cost of underusing AI.

2.1 Who we can become: enabling human self-realisation, without devaluing human skills

AI may enable self-realisation, by which we mean the ability for people to flourish in terms of their own characteristics, interests, potential abilities or skills, aspirations, and life projects. Much as inventions such as the washing machine liberated people – and particularly women – from the drudgery of domestic work, the "smart" automation of other mundane aspects of life may free up yet more time for cultural, intellectual and social pursuits or more interesting work. More AI may easily mean more human life spent more intelligently. The risk in this case is not the obsolescence of some old skills and the emergence of new ones *per se*, but the pace at which this is happening.

A very fast devaluation of old skills and hence a quick disruption of the job market and the nature of employment can be seen at the level of both the individual and society. At the level of the individual, jobs are often intimately linked to personal identity, self-esteem, and social standing, all factors that may be adversely affected by redundancy, even putting to one side the potential for severe economic harm. Furthermore, at the level of society, the deskilling of sensitive, skill-intensive domains, such as health care diagnosis or aviation, may create dangerous vulnerabilities in the event of AI malfunction or an adversarial attack. Fostering the development of AI in support of new skills, while anticipating and mitigating its impact on old ones will require both close study and potentially radical ideas, such as the proposal for some form of "universal basic income", which is growing in popularity and experimental use.

2.2 What we can do: enhancing human agency, without removing human responsibility

AI is providing a growing reservoir of "smart agency". Put at the service of human intelligence, such a resource can hugely enhance human agency. We can do more, better, and faster, thanks to the support provided by AI. In this sense of "Augmented Intelligence", AI could be compared to the impact that engines have had on our lives. The larger the number of people who will enjoy the opportunities and benefits of such a reservoir of smart agency "on tap", the better our societies will be.

Responsibility is therefore essential, in view of what sort of AI we develop, how we use it, and whether we share with everyone its advantages. Obviously, the corresponding risk is the absence of such responsibility. This may happen not just because we have the wrong socio-political framework, but also because of a "black box" mentality, according to which AI systems for decision-making are seen as being beyond human understanding and hence control. These concerns apply not only to high-profile cases, such as deaths caused by autonomous vehicles, but also to more commonplace but still significant uses, such as in automated decisions about parole or creditworthiness.

Yet the relationship between the degree and quality of agency that people enjoy and how much agency we delegate to autonomous systems is not zero-sum, either pragmatically or ethically. In fact, if developed thoughtfully, AI offers the opportunity of *improving and multiplying* the possibilities for human agency. Consider examples of "distributed morality" in human-to-human systems such as peer-to-peer lending (Floridi, 2013). Human agency may be ultimately supported, refined and expanded by the embedding of "facilitating frameworks", designed to improve the likelihood of morally good outcomes, in the set of functions that we delegate to AI systems. AI systems could, if designed effectively, amplify and strengthen shared moral systems.

2.3. What we can achieve: increasing societal capabilities, without reducing human control

Artificial intelligence offers myriad opportunities for improving and augmenting the capabilities of individuals and society at large. Whether by preventing and curing diseases or optimising transportation and logistics, the use of AI technologies presents countless possibilities for reinventing society by radically enhancing what humans are collectively capable of. More AI may support better coordination, and hence more ambitious goals. Human intelligence augmented by AI could find new solutions to old and new problems, from a fairer or more efficient distribution of resources to a more sustainable approach to consumption. Precisely because such technologies have the potential to be so powerful and disruptive, they also introduce proportionate risks. Increasingly, we may not need to be 'on the loop', if we can delegate our tasks to AI. However, if we rely on the use of AI technologies to augment our own abilities in the wrong way, we may delegate important tasks and above all decisions to autonomous systems that should remain at least partly subject to human supervision and choice. This in turn may reduce our ability to monitor the performance of these systems (by no longer being 'on the loop' either) or preventing or redressing errors or harms that arise ('post loop'). It is also possible that these potential harms may accumulate and become entrenched, as more and more functions are delegated to artificial systems. It is therefore imperative to strike a balance between pursuing the ambitious opportunities offered by AI to improve human life and what we can achieve, on the one hand, and, on the other hand, ensuring that we remain in control of these major developments and their effects.

2.4 How we can interact: cultivating societal cohesion, without eroding human self-determination

From climate change and antimicrobial resistance to nuclear proliferation and fundamentalism, global problems increasingly have high degrees of coordination complexity, meaning that they can be tackled successfully only if all stakeholders co-design and co-own the solutions. AI, with its data-intensive, algorithmic-driven solutions, can hugely help to deal with such coordination complexity, supporting more societal cohesion and collaboration. For example, efforts to tackle climate change have exposed the challenge of creating a cohesive response, both within societies and between them. The scale of this challenge is such that we may soon need to decide between engineering the climate directly and engineering society to encourage a drastic cut in harmful emissions. This latter option might be undergirded by an algorithmic system to cultivate societal cohesion. Such a system would not be imposed from the outside; it would be the result of a self-imposed choice, not unlike our choice of not buying chocolate if we need to be on a diet. "Self-

nudging" to behave in socially preferable ways is the best form of nudging. It is the outcome of human decisions and choices, but it can rely on AI solutions to be implemented. Yet the risk is that AI systems may erode human self-determination, as they may lead to unplanned and unwelcome changes in human behaviours to accommodate the routines that make automation work and people's lives easier. AI's predictive power and relentless nudging, even if unintentional, should be at the service of human self-determination and foster societal cohesion, not the undermining of human dignity or human flourishing.

Taken together, these four opportunities, and their corresponding challenges, paint a mixed picture about the impact of AI on society and the people in it. Accepting the presence of trade-offs, seizing the opportunities, while avoiding or minimising the risks head-on will improve the prospect for AI technologies to promote human dignity and flourishing. Having outlined the potential benefits to individuals and society at large of an ethically engaged approach to AI, in the next section we highlight the "dual advantage" to organisations of taking such an approach.

3. The dual advantage of an ethical approach to AI

Ensuring socially preferable outcomes of AI relies on resolving the tension between incorporating the benefits and mitigating the potential harms of AI, in short, simultaneously avoiding the misuse and underuse of these technologies. In this context, the value of an ethical approach to AI technologies comes into starker relief. Compliance with the law is merely necessary, but significantly insufficient (Floridi, 2018). With an analogy, it is the difference between playing according to the rules, and playing well, so that one may win the game. Adopting an ethical approach to AI confers what we define here as a "dual advantage". On one side, ethics enables organisations to take advantage of the social value that AI enables. This is the advantage of being able to identify and leverage new opportunities that are socially acceptable or preferable. On the other side, ethics enables organisations to anticipate and avoid or at least minimise costly mistakes. This is the advantage of prevention and mitigation of courses of action that turn out to be socially unacceptable and hence rejected. This also lowers the opportunity costs of choices not made or options not grabbed for fear of mistakes.

Ethics' dual advantage can only function in an environment of public trust and clear responsibilities more broadly. Public acceptance and adoption of AI technologies will occur only if the benefits are seen as meaningful and risks as potential, yet preventable, minimisable, or at least something against which one can be protected, through risk management (e.g. insurance) or

redressing. These attitudes will depend in turn on public engagement with the development of AI technologies, openness about how they operate, and understandable, widely accessible mechanisms of regulation and redress. The clear value to any organisation of the dual advantage of an ethical approach to AI amply justifies the expense of engagement, openness, and contestability that such an approach requires.

4. Towards a framework of principles for AI in society[1]

AI4People is not the first initiative to consider the ethical implications of AI. Many organisations have already produced statements of the values or principles that should guide the development and deployment of AI in society. Rather than conduct a similar, potentially redundant exercise here, we strive to move the dialogue forward, constructively, from principles to proposed policies, best practices, and concrete recommendations for new strategies. Such recommendations are not offered in a vacuum. But rather than generating yet another series of principles to serve as an ethical foundation for our recommendations, we offer a synthesis of existing sets of principles produced by various reputable, multi-stakeholder organisations and initiatives. A fuller explanation of the scope, selection and method of assessing these sets of principles is available in Cowls and Floridi (Forthcoming). Here, we focus on the commonalities and noteworthy differences observable across these sets of principles. The documents we assessed are:

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 Montréal Declaration for Responsible AI, developed under the auspices of the Université de Montréal, following the Forum on the Socially Responsible Development of AI of November 2017 (hereafter "Montréal"; Montréal 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);[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); and
5. the "five overarching principles for an AI code" offered in paragraph 415 of the 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).

[1] Further analysis and more information on the methodology employed in this section will be presented in Cowls and Floridi (Forthcoming).
[2] The Montréal 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 1st May, 2018.
[3] The third version of *Ethically Aligned Design* will be released in 2019 following wider public consultation.

Taken together, these five documents yield 44 principles.[4] Overall, we find an impressive degree of coherence and overlap between the five sets of principles. This can most clearly be shown by comparing the sets of principles with the set of four core principles commonly used in bioethics: beneficence, non-maleficence, autonomy, and justice. This 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). The four bioethical principles adapt surprisingly well to the fresh ethical challenges posed by artificial intelligence. But they are not exhaustive. On the basis of the comparative analysis we argue that one more, new principle is needed in addition: *explicability*, understood as incorporating both intelligibility and accountability.

4.1 Beneficence: promoting well-being, preserving dignity, and sustaining the planet

Of the four core bioethics principles, beneficence is perhaps the easiest to observe across the five sets of principles we synthesise here. The principle of creating AI technology that is beneficial to humanity is expressed in different ways, but is typically featured at the top of each list of principles. Montréal and IEEE principles both use the term "well-being"; for Montréal, "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 EGE emphasises the principle of both "human dignity" and "sustainability". Its principle of "sustainability" represents 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 these principles of beneficence firmly underline the central importance of promoting the well-being of people and the planet.

4.2 Non-maleficence: privacy, security and "capability caution"

Though "do only good" (beneficence) and "do no harm" (non-maleficence) seem logically equivalent, in both the context of bioethics and of the ethics of AI they represent distinct principles, each requiring explication. While they encourage well-being, the sharing of benefits and

[4] Of the five 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, particularly in the narrower context of academia and industry.

the advancement of the public good, each of the five sets of principles also cautions against the many potentially negative consequences of overusing or misusing AI technologies. Of particular concern is the prevention of infringements on personal privacy, which is listed as a principle in four of the five sets, and as part of the "human rights" principles in the IEEE document. In each case, privacy is characterised as being intimately linked to individual access to and control over how personal data is used.

Yet the infringement of privacy is not the only danger to be avoided in the adoption of AI. Several of the documents also emphasise the importance of avoiding the misuse of AI technologies in other ways. The Asilomar Principles are quite specific on this point, citing the threats of an AI arms race and of the recursive self-improvement of AI, as well as the need for "caution" around "upper limits on future AI capabilities". The IEEE document meanwhile cites the need to "avoid misuse", while the Montréal Declaration argues that those developing AI "should assume their responsibility by working against the risks arising from their technological innovations", echoed by the EGE's similar need for responsibility.

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. Confused also is the question of intent: promoting non-maleficence can be seen to incorporate the prevention of both accidental (what we above call "overuse") and deliberate (what we call "misuse") harms arising. In terms of the principle of non-maleficence, this need not be an either/or question: the point is simply to prevent harms arising, whether from the intent of humans or the behaviour of machines (including the unintentional nudging of human behaviour in undesirable ways).. Yet these underlying questions of agency, intent and control become knottier when we consider the next principle.

4.3 Autonomy: the power to decide (to decide)

Another classic tenet of bioethics is the principle of autonomy: the idea that individuals have a right to make decisions for themselves about the treatment they do or not receive. In a medical context, this principle of autonomy is most often impaired when patients lack the mental capacity to make decisions in their own best interests; autonomy is thus surrendered involuntarily. With AI, the situation becomes rather more complex: when we adopt AI and its smart agency, we *willingly* cede some of our decision-making power to machines. 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 principle of autonomy is explicitly stated in four of the five documents. The Montréal Declaration articulates the need for a balance between human- and machine-led decision-making, stating that "the development of AI should *promote* the autonomy of all human beings *and control* … the autonomy of computer systems" (italics added). The EGE argues that autonomous systems "must not impair [the] freedom of human beings to set their own standards and norms and be able to live according to them", 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".

These documents express a similar sentiment in slightly different ways, echoing the distinction drawn above between beneficence and non-maleficence: not only should the autonomy of humans be promoted, but also the autonomy of machines should be restricted and made intrinsically reversible, should human autonomy need to be re-established (consider the case of a pilot able to turn off the automatic pilot and regain full control of the airplane). Taken together, the central point is to protect the intrinsic value of human choice – at least for significant decisions – and, as a corollary, to contain the risk of delegating too much to machines. Therefore, what seems most important here is what we might call "meta-autonomy", or a "decide-or-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 other interests such as efficiency outweigh the loss of control over decision-making. As anticipated, any delegation should remain overridable in principle (deciding to decide again).

The decision to make or delegate decisions does not take place in a vacuum. Nor is this capacity to decide (to decide, and to decide again) necessarily distributed equally across society. The consequences of this potential disparity in autonomy is addressed in the final of the four principles inspired by bioethics.

4.4 Justice: promoting prosperity and preserving solidarity

The last of the four classic bioethics principles is justice, which is typically invoked in relation to the distribution of resources, such as new and experimental treatment options or simply the availability of conventional healthcare. Again, this bioethics principle finds clear echoes across the principles for AI that we analyse. The importance of "justice" is explicitly cited in the Montréal 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". The emphasis on the protection of social support systems may reflect geopolitics, insofar as the EGE is a European body. The AIUK report argues that citizens should be able to "flourish mentally, emotionally and economically alongside artificial intelligence".

As with the other principles already discussed, these interpretations of what justice means as an ethical principle in the context of AI are broadly similar but contain subtle distinctions. Justice variously relates to a) using AI to correct past wrongs such as eliminating discrimination; b) ensuring that the use of AI creates benefits that are shared (or at least shareable); and c) preventing the creation of *new* harms, such as the undermining of existing social structures. Notable also are the different ways in which the position of AI, *viz-à-viz* people, is characterised in relation to justice. In some cases, it is AI technologies themselves that "should benefit and empower as many people as possible" and "contribute to global justice"; in others, it is "the *development* of AI" that "should promote justice" (italics added). At still other points, people should flourish merely "alongside" AI. Our purpose here is not to split semantic hairs. The diverse ways in which the relationship between people and AI is described in these documents hints at broader confusion over AI as a manmade reservoir of "smart agency". Put simply, and to resume our bioethics analogy, are we (humans) the patient, receiving the "treatment" of AI, the doctor prescribing it? Or both? It seems that we must resolve this question before seeking to answer the question of whether the treatment will even work. This is the core justification for our identification within these documents of a new principle, one that is not drawn from bioethics.

4.5 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 our everyday life. The situation is inherently unequal: a small fraction of humanity is currently engaged in the design and development of a set of technologies that are already transforming the everyday lives of just about everyone else. This stark reality is not lost on the authors whose documents we analyse. In all, reference is made to the need to *understand* and *hold to account* the decision-making processes of AI. This principle is expressed using different terms: "transparency" in Asilomar, "accountability" in EGE, both "transparency" and "accountability" in IEEE, and "intelligibility" in AIUK. Though

described in different ways, 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 this principle, which we synthesise as "explicability" both in the epistemological sense of "intelligibility" and in the ethical sense of "accountability", is therefore the crucial missing piece of the jigsaw when we seek to apply the framework of bioethics to the ethics of AI. It complements the other principles: for AI to be beneficent and non-malificent, 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 ensure that the technology – including its human developers and deployers – are held accountable in the event of a serious, negative outcome, which would require in turn some understanding of why this outcome arose. More broadly, we must negotiate the terms of the relationship between ourselves and this transformative technology, on grounds that are readily understandable to the proverbial man or woman "on the street".

Taken together, we argue that these five principles capture the meaning of each of the 44 principles contained in the five high-profile, expert-driven documents, forming an ethical framework depicted in Figure 4.1.

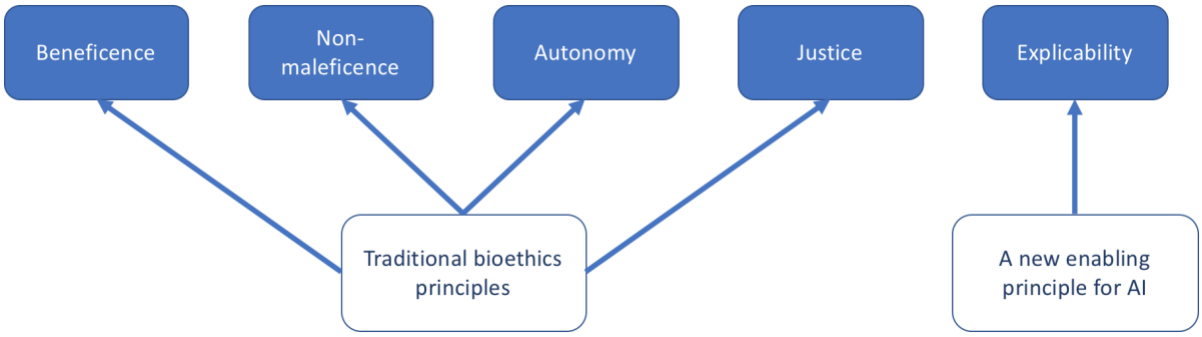

Figure 4.1: an ethical framework for AI, formed of four traditional and one new principle.

The framework of principles described and depicted above provides a set of "ethical guardrails" within which more concrete recommendations for law, policy, and best practice can be made.

Conclusion

In this paper, we have introduced what we argue are the four core opportunities for society presented by AI, as well as four corresponding risks, and the opportunity cost of underusing this set of technologies. We then explored five sets of principles provided by reputable bodies, each of which builds on evidence obtained and perspectives gathered from diverse stakeholders. The five principles which emerged (four of which can be traced directly to equivalent principles in bioethics) together offer a framework within which more concrete recommendations for law, policy, and best practice can be made.

This paper and its findings identifies noteworthy areas for future research, especially research which examines how organisations from government, industry, civil society and academia together develop concrete laws, policies and best practices for seizing the opportunities and minimising the risks posed by AI technologies. We argue that such attempts would benefit from building on ethical frameworks such as the one offered here.

References

Asilomar AI Principles (2017). Principles developed in conjunction with the 2017 Asilomar conference [Benevolent AI 2017]. Retrieved June 20, 2018, from https://futureoflife.org/ai-principles

Cowls, J. and Floridi, L. (Forthcoming). The Utility of a Principled Approach to AI Ethics.

European Group on Ethics in Science and New Technologies (2018, March). Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems. Retrieved June 20, 2018, from https://ec.europa.eu/info/news/ethics-artificial-intelligence-statement-ege-released-2018-apr-24_en.

Imperial College London (2017, Oct, 11). Written Submission to House of Lords Select Committee on Artificial Intelligence [AIC0214]. Retrieved June 20, 2018, from http://bit.ly/2yleuET

The IEEE Initiative on Ethics of Autonomous and Intelligent Systems (2017). Ethically Aligned Design, v2. Retrieved June 20, 2018, from https://ethicsinaction.ieee.org

Floridi, L. (2013). The Ethics of Information. Oxford, Oxford University Press.

Floridi, L. (2018). Soft Ethics and the Governance of the Digital. Philos. Technol. 2018, 1-8.

House of Lords Artificial Intelligence Committee (2018, April, 16). AI in the UK: ready, willing and able? Retrieved June 20, 2018, from https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm

King, T., Aggarwal, N., Taddeo, M., and Floridi, L (2018, May, 22), Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. Available at SSRN: https://ssrn.com/abstract=3183238

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. Retrieved June 20, 2018, from https://www.montrealdeclaration-responsibleai.com/the-declaration).

Taddeo, M. (2017). The limits of deterrence theory in cyberspace. Philos. Technol. 2017, 1–17.