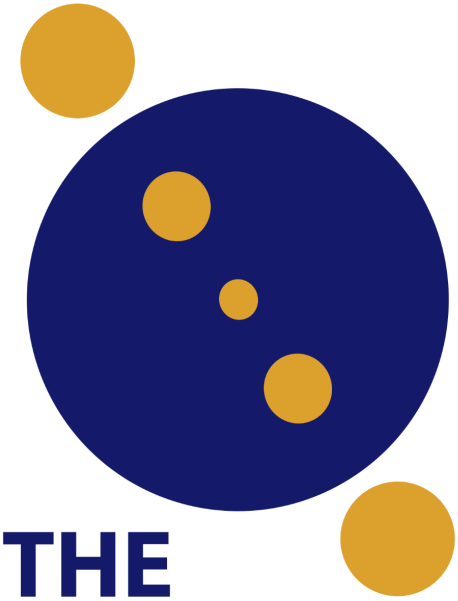

THE LYCEUM PROJECT

AI ETHICS WITH ARISTOTLE

PROFESSOR JOSIAH OBER AND PROFESSOR JOHN TASIOULAS

JUNE 17, 2024

Stanford University
Human-Centered
Artificial Intelligence

EXECUTIVE SUMMARY

Many today take the view that the AI technological revolution is creating a radically new reality, one that demands a corresponding upheaval in our ethical thinking. This can generate a sense of helplessness in the face of rapid technological advances. But the idea that we start from an ethical blank slate in addressing the challenges and opportunities of this transformative technology is a fallacy.

The key contention of this white paper is that the basic approach to ethics developed by the 4th Century BC Greek philosopher, Aristotle, offers the most compelling framework for addressing the challenges and opportunities of Artificial Intelligence today.

Among the key elements of an Aristotelian approach to AI ethics explored in the white paper are the following:

* a truly 'human centred' approach to the ethics of AI, one that conceives both human flourishing and human morality as rooted in our nature as human beings whose fulfilment depends on the exercise of capacities for rationality, social engagement, and communication. The Aristotelian approach constantly foregrounds the question of the *good* for which AI systems are developed and deployed and does not conceive of ethics as in competition with technological advance.

* a richer conception of ethics than the dominant ethical theories in the discourse of AI: on the one hand, approaches grounded in the fulfilment of preferences or the maximisation of wealth; on the other hand, approaches based on human rights law. The former are focussed on considerations that are not ultimate values; the latter are incomplete, failing to recognise that considerations besides rights, such as virtues and the common good, are essential.

* an emphasis on the powerful nexus between ethics and politics, since human beings can only flourish in communities, and the importance of democracy (Aristotle's notion of citizens as 'ruling and being ruled in turn') and liberalism (given the importance of free choice in the cultivation and exercise of the Aristotelian virtues) as political values.

* the idea of AI systems as ‘intelligent tools’ to be deployed in order to advance the flourishing of individuals and communities, a focus very different from the dominant objective of the tech industry, which is to create Artificial General Intelligence that replicates human intellectual capacities across their entire spectrum.

* an elucidation of how Aristotle’s own doomed attempt to justify the existence of a class of ‘natural slaves’ serves as a cautionary illustration of the perils of seeking to instrumentalise beings with intellectual capabilities comparable to humans.

* the development of the idea of AI systems as ‘intelligent tools’ in two domains – that of work, where their fundamental role is to augment the exercise of valuable human capacities, rather than to replace systematically human endeavour, and that of democracy, where AI tools can play an important role in enabling a genuinely deliberative and participatory democracy to operate at the scale of the modern state.

* the guideposts that the Aristotelian framework offers in regulating AI, overcoming unduly restrictive regulatory rubrics like ‘safety’ and ‘existential risk’, and providing a more comprehensive regulatory framework than the frameworks deployed by the world’s three digital empires: state-based (China), market-based (US), and rights-based (EU).

* a template for addressing the challenge of how international co-operation might be achieved in regulating AI globally without international law and institutions intruding unduly into the proper sphere of decision that belongs to states.

* a case for a novel human right for the age of AI: a qualified right to a human decision that has the effect of prohibiting certain decisions from being made by AI systems, or else allowing opt-outs or appeals from such decisions where they are permissible.

The authors gratefully acknowledge support in developing this paper from the Patrick J McGovern Foundation, the World Human Forum, the Cosmos Institute, Demokritos, Oxford University Institute for Ethics in AI, and Stanford University Human-Centred Artificial Intelligence.

INTRODUCTION

“I feel like we are nearing the end of times. We humans are losing faith in ourselves”.

These words were uttered by the acclaimed eighty-three year old Japanese animator and filmmaker, Hayao Miyazaki, in a video clip that was posted several years ago on YouTube. Earlier in the clip, Miyazaki was shown sitting across a conference table with a group of chastened-looking technologists who had just shown him an AI-generated image of a humanoid figure effortfully writhing across a floor by means of “grotesque movements, which we humans can’t imagine”. The group was seeking, as one of them explained, to “build a machine that can draw pictures like humans do”. “I am utterly disgusted”, Miyazaki rebuked them. “If you really want to make creepy stuff, you can go ahead and do it. I would never wish to incorporate this technology in my work at all. I strongly feel that this is an insult to life itself”. One of the computer scientists seemed to wipe a tear from the corner of his eye as he replied, apologetically and a little implausibly, that this was “just our experiment... We don’t mean to do anything by showing it to the world”.[1]

This episode powerfully encapsulates a conflict that goes to the spiritual crux of our current moment in the development of AI technology. On the one hand, there is the drive by technologists, technology corporations, and governments to create sophisticated AI tools that can simulate more and more of the paradigmatic manifestations of human intelligence, from composing a poem to diagnosing an illness. For many of them, the ultimate goal, at the end of this road, is Artificial General Intelligence, a form of machine intelligence that spans the entire spectrum of human cognitive capabilities. On the other hand, there is the dreadful sense that this whole enterprise, for all its efficiency gains and other supposed benefits, is an affront to our human nature and a pervasive threat to our prospects of living a genuinely valuable human life – “an insult to life itself” in Miyazaki’s words.

The conflict just described stems from the fact that human intelligence, in all its formidable reach and complexity, has long been considered the locus of the special value that inheres in all human beings. It distinguishes us from artefacts and non-human animals alike. But if machine intelligence can eventually replicate or even

[1] Manhattan Project for a Nuclear-Free World, "Hayao Miyazki's thoughts on an artificial intelligence", 2016, Available at: https://www.youtube.com/watch?v=ngZ0K3lWKRc&ab_channel=ManhattanProjectforaNuclear-FreeWorld (Accessed 15 June 2024).

out-perform human intelligence, where would that leave humans? Would the pervasive presence of AI in our lives be a negation of our humanity and an impediment to our ability to lead fulfilling human lives? Or can we incorporate intelligent machines into our lives in ways that dignify our humanity and promote our flourishing? It is *this* challenge, rather than the rather far-fetched anxiety about human extinction in a robot apocalypse, that is the most fundamental 'existential' challenge posed today by the powerful new forms of Artificial Intelligence. It is a challenge that concerns what it *means* to be human in the age of AI, rather than just one about ensuring the continued survival of humanity. Some take the view that the AI technological revolution is creating a radically new reality, one that demands a corresponding upheaval in our ethical thinking. This outlook can foster a sense of helplessness, the feeling that technological innovations are accelerating at an exponential rate, with radically transformative implications for every aspect of human life, while our ethical resources for engaging with these developments are pitifully meagre or non-existent. We reject this pessimistic and disempowering view of our ethical situation in the face of rapid technological change. We *do* already have rich ethical materials needed to engage with the challenges of the AI revolution, but to a significant degree they need to be rescued from present-day neglect, incorporated into our decision-making processes, and aced into dialogue with the dominant ideological frameworks that are currently steering the development of AI technologies - ideologies centred on the promotion of economic growth, maximising the fulfilment of subjective preferences, or complying with legal standards, such as human rights law.

Surprising as it may seem, our contention is that the basic approach to ethics developed by the 4th Century BC Greek philosopher, Aristotle, and subsequently built on by many later thinkers over the past 2,400 years, offers the most compelling framework for addressing the challenges of Artificial Intelligence today. The Aristotelian framework has a rich conception of human nature at its core, according to which we are essentially rational and social animals. It understands ethics, both in terms of what it is to live a fulfilling human life and what we morally owe to others, as essentially bound up with the exercise of distinctively human capabilities. And it is deeply political in character, identifying the essential purpose of a political community as that of enabling the flourishing of each and every one of its members. It offers a deeper and

more compelling basis for engaging with the 'ethics of AI' than the dominant frameworks.

Of course, no philosophical framework is a panacea for solving ethical problems. Indeed, Aristotle would be the first to insist on the vital need for practical wisdom that is attuned to the fine details of distinct problem situations and whose operations are not reducible to the mechanical application of pre-existing rules or theories. Moreover, on an Aristotelian approach, we must work to cultivate a cultural and institutional environment that fosters sound decision-making by individuals and communities. Nor do we endorse every specific ethical view that Aristotle propounded; indeed, some of them, such as his views on slavery and women, are grotesquely mistaken. But even these egregious errors do not invalidate the basic correctness of the general framework he elaborated.[2] And no general framework, however sound, can immunise us from human fallibility. With all these caveats in mind, the Aristotelian ethical framework can provide valuable guidance in identifying the relevant normative considerations and determining priorities among them, and it can help us to resist the dominance of influential contemporary ideas that work to make AI technologies a threat to the prospects of individual and communal flourishing. Ethics in the spirit of Aristotle, in short, is indispensable if we are to retain faith in our humanity in the age of AI. In making this case, this paper is divided into three parts.

Part I: THE ARISTOTELIAN FRAMEWORK: HUMAN NATURE, ETHICS, AND POLITICAL COMMUNITY. The Aristotelian framework for thinking about AI consists of three core interlocking ideas: (1) that human beings possess a distinctive nature as especially rational, communicative, and social animals, a nature that is not shared by existing AI systems or any such systems that are likely to be developed in the foreseeable future, and that an understanding of human nature is the basis for our ethical thought; (2) that ethics concerns the flourishing of human beings (human well-being) as individuals and members of communities, and also what they owe others, including those outside their communities and non-human beings (morality). The flourishing of each and all requires that we be free to exercise the core capacities that distinguish us as a kind of being. First is: *sociability*; we are highly interdependent, requiring social cooperation with others for our material and moral well-being. Next is *reason*, regulation of beliefs, emotions, and choices in accordance with rational judgments about both means and ends. And finally, *communication* through language and other forms of symbolic expression. Forms of social organisation (institutions,

[2] Josiah Ober, "Political Animals Revisited." *The Good Society* 22 (2013), pp. 201-214.

norms) and technology (tools and the know-how that enables their use) that advance the prosocial use of reason and communication promote flourishing; those that impede that use degrade flourishing. (3) The fundamental purpose of a political community is to secure the common (joint and several) good, i.e. to furnish the material, institutional, educational, and other conditions that enable the flourishing of each and every one of its members as free and equal citizens. The free, cooperative, prosocial use of the core capacities of reason and communication, by the diverse members of a community, is pluralistic democracy. And so, the overarching political purpose of the Aristotelian human community is not only compatible with, but requires both democracy and a form of liberalism.

Part II: AI SYSTEMS AS INTELLIGENT TOOLS. The positive vision of AI that emerges from the Aristotelian account is the idea that AI systems should be understood, developed, and deployed as 'intelligent tools' that enhance our ability to flourish as individuals and communities. They should not be regarded as means of transcending our human nature, or of creating a race of intelligent artificial beings with comparable ethical standing to humans (here Aristotle's own profoundly mistaken justification of slavery is a powerful warning). Nor should we enable them to systematically replace valuable human endeavour with machines in domains such as work, personal relations, artistic activity, politics, and so on. In the words of the American philosopher, Daniel Dennett, AI systems are "intelligent tools, not colleagues" (nor, we would add, friends, lovers, or fellow citizens).[3] This will involve characterising some of the fundamental ways in which AI systems differ from human beings with respect to capacities such as consciousness, understanding, autonomous agency, and reasoning. This part considers the idea of AI as an 'intelligent tool' in relation to three topics: (a) work and leisure, and (b) a radically participatory democratic culture.

PART III: SIGN-POSTS FOR REGULATION. The final part argues that the Aristotelian framework, and the conception of AI systems as intelligent instruments that it supports, can help steer the regulation of the development and deployment of AI, whether through domestic law, international law, 'soft' norms such as the UN Guiding Principles on Business and Human Rights or industry-wide codes of conduct. It begins by showing how the Aristotelian framework reveals the limitations of 'safety' as an overarching regulatory framework. It then proceeds to show the superiority of the Aristotelian framework to the world's three dominant regulatory approaches to digital

[3] 'Philosopher Daniel Dennett on AI, robots and religion', *Financial Times* March 3rd, 2017 https://www.ft.com/content/96187a7a-fce5-11e6-96f8-3700c5664d30

regulation - statist (China), market-driven (US), and rights-based (EU) - partly in virtue of its providing a more compelling setting within which the elements of state action, the market, and basic rights protections can be integrated. It then goes on to argue that the framework offers a basis for inter-cultural dialogue in elaborating global standards while preserving significant autonomy for distinct political communities within the architecture of global AI regulation, thereby guarding against global governance overreach. Next, it shows that Aristotle's concern with human interdependence and the conditions necessary for self-sufficiency provide the basis for an Aristotelian argument for international cooperation in the regulation of the global environment and of AI. Part III concludes by drawing on the Aristotelian framework to make a case for a novel right to a human decision as an element of the global regulation of AI

PART I

THE ARISTOTELIAN FRAMEWORK: HUMAN NATURE, ETHICS, AND POLITICAL COMMUNITY

> We're definitely wrestling with how, when we make not just grade school or middle school level intelligence, but Ph.D level intelligence and beyond, the best way to put that into a product, the best way to have a positive impact on society and people's lives. We don't know the answer to that yet. So I think that's a pretty important thing to figure out.[4]
>
> Sam Altman, Co-Founder and CEO of OpenAI

At the core of the Aristotelian ethical framework are three interlocking ideas:

(1) that human beings possess a distinctive nature as political (especially social, rational, and communicative) animals, a nature that is not shared by existing AI systems or any such systems that are likely to be developed in the foreseeable future, and that a grasp of human nature is the basis for our ethical thought;

(2) that ethics – whose subject-matter includes both the flourishing of human beings (human well-being), and what they owe others, including non-human beings (morality) – and flourishing, along with the ability to fulfil moral obligations, enjoins the free exercise of core human capacities, notably cooperative sociability, reason, and communication;

(3) that the fundamental purpose of a political community is to advance the common good of its members, i.e. to furnish the material, institutional, educational, and other conditions that enable the flourishing of each and every one of its members as free and equal citizens. The conjoined free exercise of prosocial human capacities of reason and communication leads to participatory democracy, the best form of human social organisation, the most capable of identifying and advancing common goods.

[4] Stanford eCorner, "The Possibilities of AI[Entire Talk] - Sam Altman (OpenAI)", 2024, Available at: https://www.youtube.com/watch?v=GLKoDkbS1Cg

Human nature

In the world of AI, there is much talk of the need for 'human-centred AI'. But beyond conveying a vague sense of reassurance in the face of the accelerating capacities of AI systems, what does 'human-centred' really mean? From an Aristotelian standpoint, the reference to 'human-centeredness' is no mere rhetorical ploy. Instead, an Aristotelian takes it as axiomatic that ethical inquiry requires an understanding of human nature itself. For how can we know what it is for a human life to go well or badly, or what it is we owe others, without a grasp of what kind of being a human is? For Aristotle, there is a universal human nature that unites us all and of which we can aspire to have genuine knowledge. The concept of the 'human' is not a mere social construct that is subject to radical variation from one society to another. To a large extent, knowledge of human nature will be produced by inquiry in the natural sciences, most notably biology, because humans are animals of a certain kind, adapted, as is the case of all vertebrate animals, to live in a natural environment, and possessing features such as embodiment, mortality, and basic needs for air, food, water, sex, and shelter.

But the natural sciences are not the only source of objective truths about human nature. Humans resemble other political animals (Aristotle's examples are bees and ants) in their dependence on collectively produced common goods. But it is a feature of human beings that they express their own self-understanding in highly complex cultural patterns of social life, in laws and institutions, historical narratives, artistic creations, religious practices, and so on. Understanding human nature also demands that we take these self-conceptions seriously, drawing critically not only on common opinion (*endoxa*), but also on visions of humanity elaborated in art, history, literature, music, and philosophy. This methodological pluralism is related to the fact that the study of human nature is not a value-free inquiry but rather inextricably intertwined with understanding what it is for things to go well or badly for beings such as us. Because Aristotle has this expanded sense of the 'natural', one which includes evaluative elements, he is not prey to the charge of committing the Humean fallacy of deriving claims about what 'ought' to be from statements about what 'is' the case. More generally, reliance on Aristotle's evaluative conception of human nature does not involve an outmoded metaphysical teleology that is incompatible with modern science.[5] Finally, of course, our characterization of human nature is not immutable, but provisional and fallible, open to ongoing revision in light of new experiences, especially

[5] Martha C Nussbaum, "Aristotle on Human Nature and the Foundations of Ethics" in J.E.J. Altham and Ross Harrison (eds), *World, Mind, and Ethics* (Cambridge University Press, 1995), pp. 86-131.

those produced by engagement with the ideas and practices of people from different times and places than our own, since they too share a common human nature with us.

Humans are, for Aristotle, the *most* political of animals because of our distinctive capacities that enable us to produce the various aspects of culture sketched above. Aristotle puts special emphasis on three aspects of human nature. Like other political animals, we produce goods that are essential to the survival and well-being of the community and of the individuals that comprise it. Living well outside a community would signal one is less or more than human: beast or god.[6] We are the most political of animals because we produce not only material common goods (food, shelter), but moral goods: our flourishing requires a just community in which the full array of human excellences (virtues) can be developed and manifest. This is only possible because humans, uniquely among animals, have the capacity to reason, not only about advantage and disadvantage, but also about right and wrong, good and evil. We deliberate about such things, not only internally, through private contemplation, but through active symbolic communication with others - through language. Our human nature thus involves capacities necessary for human morality.

In taking seriously the rootedness of human morality in human nature, the Aristotelian framework is at odds with some trends of thought that have been prominent in the world of AI. Among these is the project of creating Artificial General Intelligence (AGI), a form of AI that replicates human intellectual capabilities across the board, including responsiveness to moral concerns. If human morality is keyed to essential features of human nature, then it is hard to see how any being that did not share our human nature would be appropriately either subject or attuned to such a morality. Instructive here is the example of non-human animals and the Olympian gods whose behaviour is often morally grotesque when judged according to the standards of human morality. An objectively different nature entails the applicability of an objectively different morality.

Another project that is called into question is that of using AI and related technologies to transcend our human nature, for example, by evolving into cyborgs or 'uploading' ourselves onto the cloud or living in virtual reality. Instead, the focus is on the familiar natural world into which we were born, where we encounter plants, rivers, mountains, the sky, and people and animals in their original biological form, all locked into complex forms of interdependence. One problem here is that being human, in a biological

[6] Aristotle assumed a ranked moral hierarchy, from non-human animals, to humans, to gods and also assumes (*Politics* 1256b15-22) that plants and animals exist for the sake of human welfare. We accept none of these premises, nor are they necessary (or helpful) for our Aristotelian argument here.

sense, is an essential aspect of our identity as individuals, so that overcoming our human nature would effectively mean ceasing to exist as the distinct and interdependent individual that each of us is. Moreover, insofar as such proposals do not simply offer a deeply impoverished form of life, they involve a vast leap into the dark, since the abnegation of our humanity and the radical detachment from our natural environment that they entail confronts us with deep uncertainty of the ethical shape of the world we would inhabit after we had made the leap into trans-humanism or virtual reality.[7] Finally, the Aristotelian framework is also opposed to more moderate and more widespread forms of scepticism about human nature, such as the idea that human beings are irredeemably irrational, and also utterly opaque in their decision-making processes, and that therefore we should strive as far as possible to replace consequential human decision-making with decision-making by machines.[8]

Ethics

We here highlight three major themes in the Aristotelian account of ethics, the importance of choice, the richness of ethics, and the substantive and uncodifiable nature of practical reason.

The importance of choice. On the Aristotelian view, ethics is a domain of individual and collective human choices based on reason. As rational animals, capable of reasoning about ends as well as means, we have the capacity to stand back from our present desires, or from socially established patterns of life, and to ask what it is we should do in light of the reasons for and against any course of action. This kind of deliberation presupposes the reality of human choice from among a plurality of options. "No one", says Aristotle, "deliberates about things that are invariable, nor about things that it is impossible for him to do" (*Nicomachean Ethics* [=*NE*] VI.5).

Too often today, however, influential voices present the development of AI-based technologies and their increasing penetration into all domains of our life as inexorable processes over which we can exercise little or no control. One recent book uses the metaphor of the rise of AI as akin to a tidal wave that is hurtling towards us, something

[7] For questions like these about living in virtual reality, see David Chalmers, *Reality+: Virtual worlds and the problems of philosophy* (W.W. Norton & Company, 2022).
[8] For a perceptive discussion of the denigration of human nature that is part and parcel of tendencies described above, see Meghan O'Gieblyn, *God Human Animal Machine: Technology, Metaphor, and the Search for Meaning* (New York City, Doubleday, 2021).

we cannot fundamentally affect. [9] As with the rhetoric of inevitability that previously surrounded the topic of economic globalisation, the denial that there are choices often serves to mask both the fact of their existence and the identity of those making them. The reality is that choices are made, but in the service of the self-interest of those making them. In this case, the end being sought is, in Aristotle's terms, unjust: it is the enrichment of a part (the powerful few) at the expense of the common good of the whole (the wider community). This sort of technological determinism is, moreover, anti-democratic: massively disempowering for individuals and political communities alike. We need to understand, following Aristotle, that ethics is a domain in which we deliberate about what to do on the basis that we have effective choices over: ends *and* the means to those ends - and politics is the domain in which we make those choices and act on them, together. The exercise of a capacity for choice is inherent to our dignity as social, rational, communicative animals, which is what Aristotle tells us we most fundamentally are.

The emphasis on human choice needs to be pressed both against optimistic and pessimistic versions of the techno-determinist narrative. For example, against the comforting notion that technological progress generates broad-based prosperity through some inexorable process, Daron Acemoglu and Simon Johnson have argued that

> shared prosperity emerged because, and only when, the direction of technological advances and society's approach to dividing the gains were pushed away from arrangements that primarily served a narrow elite…[A] thousand years of history and contemporary evidence make one thing abundantly clear: there is nothing automatic about new technologies bringing widespread prosperity. Whether they do or not is an economic, social, and political choice.[10]

And what applies to technology bringing prosperity also applies to its bringing deprivation and disaster. Human choices are central either way.

---

[9] Mustafa Suyleman, *The Coming Wave: AI, Power and the 21st Century's Greatest Dilemma* (Bodley Head, 2023), p.6. And elsewhere in the book: 'Exponential change is coming. It is inevitable. That fact needs to be addressed', p.225.

[10] Daron Acemoglu and Simon Johnson, *Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity* (Basic Books, 2023), pp. 6, 13. The whole book is a powerful critique of the 'productivity bandwagon' thesis that new machines and production methods that increase productivity inexorably generate widespread prosperity by increasing the demand for workers which in turn leads to higher wages.

The richness of ethics. A second lesson to be learnt from Aristotle concerns the richness of the ethical considerations that bear on our choices. His view of ethics goes back to Socrates' question in Plato's *Republic*: How should one live? In addressing that question, we have to be attentive to two further questions and their inter-relations: 1) what makes for a flourishing life (well-being)?, and 2) what do we morally owe others – other humans, other animals, or nature itself (morality)? Moreover, an Aristotelian conception of the goal of an ethical life – *eudaimonia,* 'living well and doing well' – regards well-being and morality as deeply intertwined, since a flourishing life is one that centrally involves "an activity of the soul in accordance with reason" and with virtue, in the course of a complete life (*NE* 1098a7-18). This encompasses the cultivation of virtues of character that are moral in nature, such as justice, courage, temperance. For Aristotle, to live flourishing lives in flourishing communities, requires creating and sustaining the conditions (material and educational) for the development of virtuous dispositions, such that doing the right thing, for the right reason, at the right time, becomes habitual. The upshot is that virtuous activity - the manifestation of human excellence in everyday lives and public institutions - can become normal and expected. Accordingly, an Aristotelian ethics helps us overcome the modernist idea that acting morally and acting in pursuit of one's own true self-interest are radically distinct and potentially systematically conflicting pursuits.

On the best interpretation of the Aristotelian framework, we believe that the answer to both the question of well-being and that of morality is highly pluralistic in character. While all seek the common good, that good is not comprehensive or all-encompassing: indeed pluralism is a feature, not a bug, of Aristotelian democracy (Section II). Neither well-being nor morality reduces to one master value that is to be optimised. Instead, there is an irreducible diversity of goods that can feature in a life of well-being, such as knowledge, friendship, achievement, pleasure and so on. Equally, moral demands are many and diverse, such as justice, courage, charity, loyalty, and so on. This pluralism means not only that judgement and trade-offs are inevitable in responding to the clash of goods in particular situations, but that often there may not be a single 'optimal' way to respond to the diverse value considerations that are in play, but a range of eligible options. Here, Aristotle's own tendency to prioritise a life centred on intellectual contemplation is to be resisted as insufficiently attentive to the value pluralism his own theory assumes (see the discussion of practical reason, below).

In contrast to this rich conception of ethics, a hollowed-out notion of 'ethics' is often in play in discussions about AI ethics. Most notoriously, the tech industry has sought to

equate ethics with self-regulation and the absence of legally enforceable regulations. But the reduction of ethics to self-regulation is a travesty from the Aristotelian point of view, which holds that law is an essential part of the educational, as well as correctional, apparatus of a flourishing community. If ethics is about what it means to live a good life, and what we owe to each other, then it is fundamental to *all* forms of AI regulation – from my self-regulation in deciding whether to buy a social robot to keep my elderly mother company to legally enforceable rules prohibiting the use of AI for facial recognition or social credit. Ethics is not one form of regulation among others. This is why Aristotle's *Politics* seamlessly follows on from his *Nicomachean Ethics* – it is impossible to do ethics properly without considering how we flourish as members of political communities and what we owe to our fellow citizens; man is by nature a "political animal", who can only flourish in community with others.

Another way ethics has been diminished is that it is often conceived negatively, as a series of restrictions on technological progress. In a recent lecture delivered at the University of Oxford, the founder of DeepMind, Demis Hassabis, spoke about the benefits that his AlphaFold system could bring about by massively accelerating the process of predicting the 3D structure of proteins.[11] These predictions have the potential to help in the pursuit of such valuable ends as developing malaria vaccines, protecting honey-bees, and mitigating the effects of plastic waste. Towards the end of the lecture, Hassabis said he would at last address ethical issues, such as concerns about privacy. But, of course, the lecture was ethical from the very beginning; after all, scientific understanding, health and the protection of nature are among the great goods of human life, and hence a central part of the ethical. We must reject the fallacy that AI technology confronts us with a trade-off between 'ethics' and 'technical progress'. From an Aristotelian point of view, the very progress that AI should seek to bring about – such as enhanced health care, scientific understanding, or access to justice – is already itself a matter of ethical values, not something to be contrasted with them.

Once this is grasped, it becomes obvious that we need to articulate what exactly are the benefits that the development and deployment of AI systems in any instance promises to secure. One of the greatest failures in contemporary AI regulation is an excessively narrow template for assessing the potential benefits of AI. When benefits are talked about, they often take the form of economic growth or innovation, as in the

[11] University of Oxford, "Dr Demis Hassabis: Using AI to Accelerate Scientific Discovery", 2022. Available at: https://www.youtube.com/watch?v=AU6HuhrC65k (Accessed 15 June 2024).

UK's White Paper on AI.[12] Yet neither economic growth nor innovation are themselves ultimate values. Many things could exemplify technical innovation or promote economic growth, from development of weapons of mass destruction to the sale of addictive narcotics. Innovation and growth are at best very imperfect proxies for genuine values and at worst slogans invoked to advance the wealth and power of some at the expense of others. This is especially so when the objective of wealth-maximisation is stressed without attention to how that wealth is to be justly distributed. The result is that AI technology risks driving a form of what Acemoglu and Johnson call "so-so automation" that systematically replaces human workers but generates only marginal productivity gains, thereby aggravating existing economic and status inequalities.[13]

A closely related approach to assessing the benefits of AI, which has found great popularity in recent years, is some form or other of Benthamite utilitarianism, named after the 18th Century philosopher Jeremy Bentham.[14] Utilitarianism seeks to reduce ethics to a single optimising principle: the morally right thing to do is that which will optimise the aggregate welfare, understood as the greater balance of pleasure over pain or the preference-satisfaction of all. Its understanding of welfare is data-driven: it turns on what will in fact give people pleasure or satisfy their preferences. And even if not strictly speaking an algorithm, utilitarianism purports to be a 'felicific calculus' that minimises the need for human judgement in determining what one ought to do. The enduring appeal of Benthamite utilitarianism is not hard to grasp. Intellectually, it basks in the reflected glow of science, the source of the most spectacular and consequential technological achievements in modern times. Morally, it seems egalitarian: it takes data about everyone's happiness or preferences into account, counting everyone's welfare equally. And by minimising the need for 'judgement' it curtails the risk of what Bentham called "sinister interests" biasing the impartial assessment of the general welfare.[15]

Given the methodological affinities between utilitarianism and machine learning-based AI, it is unsurprising that utilitarianism has acquired a strong following in the AI community. This is especially the case in the issue of 'aligning' AI with our values. We

[12] The White Paper conceives of a 'proportionate' approach to regulation as balancing innovation and economic growth against various risks regarding safety, fairness, etc. Yet neither economic growth nor innovation are themselves ultimate values to be set against concerns such as fairness.
https://www.gov.uk/government/publications/ai-regulation-aproinnovationapproach/whitepaper#:~:text=Pro%2Dinnovation%3A%20enabling%20rather%20than,promote%20and%20encourage%20its%20uptake (March 29, 2023).
[13] Acemoglu and Johnson, *Power and Progress*, ch. 9.
[14] On the algorithmic pretensions of utilitarianism, see Onora O'Neill, *From Principles to Practice: Normativity and Judgment in Ethics and Politics* (Cambridge University Press, 2018), pp. 59-60, 167-9.
[15] Jeremy Bentham, *A Fragment on Government* (Cambridge University Press, 1988[1776]), p.59.

see this, for example, in the recent book *Human Compatible*, by one of the world's leading AI scientists, Stuart Russell. Russell addresses the problem of ensuring that AI-based technology does not spiral out of control, unconstrained by human morality. But he uncritically assumes that human morality consists in optimising the satisfaction of human preferences.[16] But the pull of utilitarianism is stronger still in our culture, reaching beyond the tech world and academia to policy-makers and governments, partly through its historical influence on the discipline of economics. The overwhelming emphasis on economic growth, which we mentioned above, can be seen as effectively positing wealth-maximisation as the more readily measurable proxy for either pleasure or preference-satisfaction.

Why is the Benthamite approach a hollowed out conception of ethics? To begin with, well-being does not reduce to subjective experiences. Pleasure matters, but so does acquiring understanding, valuable friendships, and achievement. Similarly, preferences may be ill-informed by the facts or skewed by prejudices of various sorts or the outgrowth of subjection to oppressive practices. Equally, there are serious challenges confronting the idea that what we morally ought to do is maximise overall well-being. From an Aristotelian standpoint, we need an ethics that is tailored to the human condition. The utilitarian idea that we have the ability to survey all the options available to us, to calculate which one will maximise overall well-being, and to act on the basis of that calculation, is a double fantasy. It flies in the face of our limited cognitive capacities and our limited capacity to sacrifice our personal interests to the impartial maximisation of welfare.

But, perhaps more fundamentally still, utilitarianism creates the grotesque prospect of sacrificing the vital interests and rights of those who are losers in the process of welfare aggregation. If enough Romans derive enough pleasure from the spectacle of a small number of Christians being fed to the lions, then on utilitarian calculations, feeding them to the lions may not only be permissible, it may be morally required. The Aristotelian will agree, of course, that we need to operate with a conception of the common good, especially in political decision-making, but as we show below, it is one that differs profoundly from the utilitarian measurement of an aggregated and inherently subjective good.

The substantiveness and uncodifiability of practical reason. Core to Aristotelian ethics is the idea of practical reason, or *phronesis*, that enables us to discover truths about how

[16] Stuart Russell, *Human Compatible: Human Compatible: AI and the Problem of Control* (Allen Lane, 2019), p.178.

to live. If ethics were simply about subjective opinions or established cultural practices, it would ultimately reduce to a power struggle when peoples' ethical views clash. The Aristotelian tradition offers a more hopeful perspective. We can engage our rational powers, especially in active dialogue with others, to discover what makes life worth living and what we owe to others. Ethics is not simply a matter of subjective preferences or entrenched cultural assumptions, it pursues objective truths. It follows that in the pursuit of the good, whether in our personal choices, or in collaboration with others in our families, schools, workplaces, or the broader political sphere, we can engage in a process governed in large part by the rational pursuit of objective truth and the virtues of honesty, humility, and respectful dialogue that this pursuit demands.

By contrast, many within the world of AI adopt an impoverished conception of intelligence. This focuses on value-free means-end reasoning. Aristotle's ethics was developed in contradistinction to the strategic reasoning taught by the Sophists. According to the Sophists (and their modern analogues), rationality is all about effective means. The Sophist taught his students the *techne* (craft or art) of strategic calculation; the goal was getting one's desired end, whatever that was, using whatever means (often sophisticated rhetoric) would most efficiently achieve that goal. On this view, the question of the value of the ends and the moral appropriateness of the means are treated as matters extrinsic to the operations of intelligence. Even a serial killer, on this understanding, can exhibit flawless intelligence. Hence the worry that exercises thinkers like Russell, that a supposedly "Superintelligent" AI will be too morally obtuse to realise that it shouldn't exterminate humanity if this turned out to be the most efficient way of achieving its goal of increasing the production of paper clips.[17] Aristotle called that sort of intelligence "cleverness" (*deinotes NE* 1147a24-28). He recognised it as an essential, but subsidiary part of practical reason. As he clearly saw, ethics requires a richer conception of intelligence, one that includes the evaluation of goals and of the morally appropriate means of pursuing them.[18]

A related point is that much of the discourse of AI is about replacing human decision-making with algorithmic systems that will be more efficient and free from human biases. Against this tendency, we need to rediscover Aristotle's idea that practical wisdom cannot be reduced to the mechanical application of rules, which is what an algorithm involves. Even the best rules we can devise, says Aristotle, will encounter unforeseen circumstances in which their rigid application would lead to bad

[17] Stuart Russell, *Human Compatible*, p.167.
[18] Josiah Ober, *The Greeks and the Rational* (University of California Press, 2022), pp. 380-383.

or even disastrous consequences (*NE* V.10). We therefore need to understand that there is an ineliminable role of judgement, and an ineliminable imprecision, in ethical matters. For all the allure of a mathematical ideal of precision held out by moral theories such as utilitarianism, Aristotle reminds us that we should only aspire to as much precision as a given type of subject-matter permits (*NE* I.3). And in ethical matters, this sort of mathematical precision is out of place given severely limited susceptibility of values to quantification and the great variability in the problem situations that confront us.

If we take seriously not only the objectivity of ethical values, but also their plurality, then we will also be led to see that often there is no single correct answer to a given ethical question, but rather a bounded range of equally acceptable answers. Objectivity does not imply singularity. This is because there is more than one rationally acceptable ordering of the values in question. For example, in choosing among different job candidates, or in the sentencing of criminals, the operative values in a given instance, e.g. enthusiasm, expertise, loyalty, honesty, etc in the former case, retributive justice, mercy, deterrence, in the latter, may be acceptably balanced against each other in different ways. Contrary to the utilitarian, there is not some uniform quantitative scale on which all our values can be arrayed so that we can identify the optimal decision in each case. This means that practical reason may only take us so far in decision making, and that at some point a choice from eligible alternatives is needed. Different individuals and communities will make different choices in similar circumstances, thus defining their own particular path through life, and forming their characters and traditions accordingly.

Political community

According to the Aristotelian framework, far from being apolitical, ethics deeply informs the fundamental objective of a political community, which is to furnish the material, institutional, educational, and other conditions that enable the flourishing of each and every one of its members as free and equal citizens (‘the common good’). This follows from the fact that humans, as social, reasoning, and communicative beings, are by nature ‘political animals'. As humans, we will fail to flourish absent the opportunity to freely exercise our constitutive capacities. Active membership of a political community or *polis*, is what makes that possible, and thus enables human flourishing. To be capable of flourishing outside of any community, says Aristotle, one must be either beast or a god, not a human (*Politics* 1253a28-29)

The claim that political communities must be oriented to advancing the common good of their citizens may sound commonsensical, but the Aristotelian interpretation of this idea is highly controversial in contemporary Western political philosophy. On the one hand, it opposes liberal theories that require the state to be neutral with respect to conceptions of human flourishing, just as it should be neutral with respect to matters of religious doctrine. On the other hand, there are utilitarian theories that agree that the promotion of the common good is the correct objective of state policy, but they give the common good a radically subjectivist and maximising interpretation that is incompatible with the Aristotelian framework.

Abstracting from these philosophical disagreements, the key question that confronts an Aristotelian approach to politics is whether it can defend a liberal democratic form of government as the best, or at least one of the best, regime types in the contemporary world. As we have already suggested, above, Aristotle's naturalism indeed allows for that defence. But it is important to take democracy and liberalism separately, because democracy is conceptually distinct from liberalism, even if one believes, as we do, that the best sort of political regime in contemporary conditions is a liberal democratic one.[19]

Democracy

One of the most urgent questions is how to subject the development and deployment of AI technology to democratic control in order to ensure that it is directed to genuine goods and that the benefits it yields are fairly distributed.[20] But can we really present the Aristotelian framework as a basis for democratic governance of AI, since like many philosophers of his time Aristotle seemed to be a sceptic about democracy? Our contention is that not only is the Aristotelian framework *compatible* with a robust commitment to democracy, but that it actually demands *a more radically participatory* democratic ideal than is currently exemplified by the world's leading democratic states. To the extent that this ideal can be realised in contemporary conditions, it promises to

[19] Josiah Ober, *Demopolis*. For an impressive and wide-ranging discussion that also seeks to elaborate a liberal and democratic political vision on the basis of Aristotelian foundations see Martha C. Nussbaum, 'Aristotelian Social Democracy' in R. B. Douglass, G.M. Mara, and H.S. Richardson (eds), *Liberalism and the Good* (Routledge, 1990), pp. 203-252.

[20] This is the theme of Daron Acemoglu and Simon Johnson, *Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity* (Basic Books, 2023).

address the crisis of confidence that currently afflicts even long-established democratic states throughout the world.[21]

We begin with the pragmatic point, that it was democratic ancient Athens that witnessed the greatest flowering of philosophy in human history. Philosophers like Aristotle voted with their feet in choosing to live in a democracy, whatever their theoretical reservations. Aristotle believed that democracy was the best of the three-commonly-existing regime-types of his day: better, that is, than oligarchy or tyranny (*Politics* 1289b2-5). But, like them, he worried that democracies of his day were characterised by the unjust rule of a part of the polis' population, in its own interest over the whole: In the case of democracy the poor ruled, in an oligarchy, the rich (*Politics* III.8). He also excluded the majority of the population from active citizenship, predicated on his grossly mistaken views that women were not properly equipped to engage in rational deliberation about the common good and that there is a class of human beings who are 'slaves by nature' and hence can justly be used as means to the ends of others.

More positively, Aristotle's philosophical method was democratic in its assumption that in general humans can trust that the exercise of their rational capacities will give them a reliable picture of the world and of the goals they should pursue. Hence the starting-point of philosophical discussion for Aristotle is always the *endoxa*, the widely held opinions on a given topic. This starting point, along with Aristotle's conception of humans as political animals, provides ample material for an Aristotelian theory of democracy as collective self-rule by free and equal citizens.[22] Self-rule involves citizens' collective deliberation and decision-making with respect to the pursuit of the common good of the community. This is not simply a matter of aggregating citizens' preferences but their deliberating on what will lead to the flourishing of each and every member of the political community and making decisions accordingly. Aristotle tells us that political (as opposed to despotic) government is 'government of free and equal citizens' (*Politics* 1255b20) and he conceives of the citizen, in the first instance, as the citizen of a democracy (*Politics* 1275b5). He repeatedly asserts that in a community of moral equals, government should be participatory: by citizens 'ruling and being ruled in turn'. This conception of political rule cannot be restricted to voting for representatives

[21] For instance, an international study revealed that 42% of the youngest generation (18-35 years old) find forms of authoritarian government preferable to democracy. For the full report, see Open Society Foundations, "Open Society Barometer: Can Democracy Deliver?", 2023, pp.1-46.

[22] Josiah Ober, "Nature, History, and Aristotle's Best Possible Regime." Pp. in *Aristotle's 'Politics': A Critical Guide*, edited by T. Lockwood and T. Samaras (Cambridge: Cambridge University Press,) 224-243

in elections every three or four years. On this definition, most members of contemporary democracies are not truly citizens, because their role is too passive; they live, in reality, in a modern-day oligarchy. To this extent, Aristotle is a more radical democrat than contemporary philosophers of democracy, such as John Rawls and Jurgen Habermas, who operate with a fundamentally representative form of democracy.

An equally important line of argument appeals to the intrinsic value of engaging in the enterprise of collective self-government without a boss. That is, a form of political organisation that pays proper respect to the freedom and equality of all members of the political community, which does not infantilize them by having political choices pre-filtered by a group of experts, such as supreme court judges or central bankers. This is not to deny that some citizens are more intelligent or virtuous or have greater subject-matter expertise than others; rather, it is to insist that there is a threshold level of rational capacity to which all those who are eligible to be citizens have attained, and a proper respect for that capacity consists in the democratisation of political deliberation and decision-making. It is this second line of argument that plays an important role in conferring legitimacy upon democratic decisions even in those cases in which we have reason to believe that the outcome is sub-optimal, and perhaps even in various ways unjust.

Moreover, we need to recognise that the site of democratic deliberation is not confined to formal political institutions but extends to the culture as a whole, the *agora* not just the *ekklesia*. This importantly includes the processes through which the large tech corporations are governed, not just because a small number of companies wield massive economic and political power, but also because they often do so through undertaking what looks like a governance function (e.g. content moderation on social media platforms).

So, we conclude that the Aristotelian framework, rather than being anti-democratic, upholds a form of democracy that is so radically participatory as to invite the objection that it is inapplicable to modern states, given their immense size and the profoundly heterogeneous nature of their populations.[23] How can citizens of such societies feasibly engage in ruling as well as being ruled in turn? As we go on to discuss in Part II, one of the most conspicuous benefits of AI and digital technology is that they can potentially

[23] The problem of scale, with special reference to democracy: Brook Manville and Josiah Ober, *The Civic Bargain* (Princeton University Press, 2023).

help us address this problem of scale, to find imaginative new ways of enabling meaningful democratic participation. It could be, therefore, that AI technology is a tool that helps us transcend the crisis-ridden, eighteenth-century model of representative democratic government, in favour of a more radically participatory form of citizen self-rule that is more in keeping with the spirit of the Aristotelian framework.

Liberalism

The meaning of 'liberalism', like that of 'democracy', is heavily contested. But minimally it is a political outlook that places great emphasis on freedom – conceived of as the ability to autonomously reach decisions regarding how one should live and the liberty to pursue those decisions – as a core political value, one that plays an important role in shaping both the objectives of the political community and the constraints it should adhere to in pursuing its objectives. As noted above, some modern versions of liberalism, such as that defended by John Rawls, insist that the state must be neutral with respect to 'comprehensive' conceptions of the human good. But there are other historically prominent variants forms of liberalism – such as those elaborated by John Stuart Mill, T.H. Green and Joseph Raz – that take the object of state power to be the promotion of human flourishing, and regard freedom as an important component of such flourishing, not least in the context of modern pluralistic societies. It is with the latter sort of liberalism that we believe the Aristotelian framework – despite Aristotle's own illiberal views on specific topics – can be squared.

To begin with, many of the protections that a liberal state must afford will overlap with those that must be in place for a robust democracy to exist. Democracy, as a valuable form of collective self-government, should not be confused with a crude majoritarianism: per Aristotle's concerns about democracy in his own day, that risks unjust rule by a self-interested part over the whole. Instead, it requires norms ensuring the freedom and equality of all citizens is properly respected: it is the *free* exercise of our natural prosocial capacities that enables true flourishing. For the freedom of exercise of reason and communication to be real, there must be norms constraining individuals, organisations, and the state itself, from threatening it. These norms, which can be constitutionally entrenched against being readily overturned by democratic decision, rule out such things as the censorship of political views or gross inequalities in economic power which erode the ability of citizens to contribute to political deliberation and decision-making on a suitably equal footing. Democracy, as collective

self-government, is necessary, but insufficient for flourishing. Since there can be illiberal democracies, the demands of liberalism as free exercise sometimes outrun those of democracy as self-government.

One way to understand those additional demands is in terms of the doctrine of human rights, as broadly represented by the Universal Declaration of Human Rights. Although there have been some prominent modern Aristotelian philosophers, such as Alasdair MacIntyre, who have written that the idea of natural or human rights – moral rights inhering in all human beings simply in virtue of their humanity – is part of an 'Enlightenment project' incompatible with an Aristotelian approach to ethics, we believe they are mistaken.[24] If an 'Enlightenment' approach to human rights seeks to ground them without reference to a rich account of human flourishing, then it is profoundly misguided. But there is nothing in the idea of human rights that insists that they be grounded in this way. On the contrary, many of the leading accounts of human rights seek to ground them in universal human interests – universal elements of what it is to flourish as a human being.[25] Human rights are best seen as evolving over time in line with technological and other changes in our capacity to meet the moral demands that the basic interests of all human beings place on us, and, in Part III we address the issue of new human rights in the age of AI.

An additional point is that liberalism, as a general political and cultural outlook, also embraces the promotion of human freedom in ways that go beyond anything that can be demanded as a matter of human rights or even rights more generally. For example, it can be part of the temperament of a liberal-minded individual to exhibit toleration towards the free choices of others, however eccentric or otherwise problematic, that goes beyond anything those others have a right to insist on. Equally, a liberal society will seek to cultivate the common good of environments that support experimentation with unconventional ideas, lifestyles, modes of association and so on in ways that do not simply reduce to upholding the rights of all involved. Here, as elsewhere, virtue exceeds respect for rights.

In reaching these wider liberal conclusions from Aristotelian premises, a number of basic ideas play a key role. The first is the emphasis on choice in the cultivation of a

[24] Alasdair MacIntyre, *After Virtue*: *A Study in Moral Theory* (Bloomsbury, 2013[1981]), ch. 5.
[25] James Griffin, *On Human Rights* (Oxford University Press, 2008); Amartya Sen, *The Idea of Justice* (Harvard UP, 2009); Joseph Raz, 'Human Rights without Foundations', in S. Besson and J. Tasioulas (eds), *The Philosophy of International Law* (OUP, 2010); John Tasioulas, 'On the Foundations of Human Rights', in M. Liao et al (eds), *Philosophical Foundations of Human Rights* (Oxford University Press, 2015), pp. 45-70.

virtuous character. The virtuous person is someone who has cultivated the settled disposition freely to choose ethically sound options and to do so for the reasons that make those options sound (unlike the merely self-controlled but not virtuous individual who may do the right thing, but not for the reasons that make it right). Coerced choices cannot count as virtuous, which places severe limits on the ability of the state to promote virtue. Second, it will seldom be the case (as discussed above) that there is a single correct option in each case; rather, there is likely to be a range of options that are permissible, each representing a variety of eligible trade-offs of the different values in place. The considerable room for discretion that value pluralism creates means that individual good lives can take remarkably different forms (e.g. the life of a research scientist versus the life of a care worker) just as, at the collective level, different societies may arrive at notably different specifications of the common good (e.g. a more pastoral society versus a more technologically-oriented one).

In short, a recognisably liberal political outlook can emerge from the Aristotelian framework provided that autonomy and liberty are given their proper due as vital dimensions of a flourishing life. Of course, Aristotle himself advocated a number of notably illiberal policies, such as compulsory abortion and infanticide as means of population control and the imposition of strict age restrictions on marriage and procreation.[26] But human beings, by Aristotle's own reckoning, are rational animals. It is highly unlikely that a flawless account of the human good, complete in every detail from theoretical foundations to practical prescriptions, was arrived at by Aristotle in the 4th Century BC without any need for significant revision or supplementation in light of human experience over the course of the intervening centuries. On the contrary, one of the great insights of ethical thought since the time of Aristotle has been a more vivid appreciation of the value and demands of individual freedom. The mistake has been to assume that this insight needs to be elaborated outside of, or in opposition to, the Aristotelian framework rather than being integrated within it to the mutual advantage of both the idea of freedom and the Aristotelian framework.

[26] *Politics* VII.16 on marriage and pregnancy; VII.17 and VIII on regulation of physical and cultural education.

PART II

AI SYSTEMS AS INTELLIGENT TOOLS

Within the Aristotelian framework, AI is first and foremost a technology, one with potential for both harmful and beneficial uses, which needs to be responsibly incorporated into our lives as an enabler of individual and collective human flourishing.

A lot of the potential impact of AI on our lives is harmful: it is well-attested that AI systems are liable to deeply troubling forms of bias and discrimination in their operations, along the lines of race, sex, class, and so on;[27] their black box character threatens to render opaque to us the explanation for the various highly consequential impacts they have on our lives;[28] their dependence on vast stores of personal data for the training of algorithms raises serious issues relating to privacy and intellectual property;[29] they have a very significant environmental impact, both in the training of such systems and in their subsequent use;[30] they threaten systematically to displace human workers from their jobs and, more generally, they can have unintended side-effects in eroding the capabilities of those individuals and communities that become over-reliant upon them;[31] they might be weaponised by 'bad actors', whether it is those spreading disinformation and misinformation that subverts democracy, or those who employ them on social media platforms in order to maximise the amount of time people spend on them, often by promoting extremist views and stoking political polarisation, or by governments or corporations using them as mechanisms of surveillance and control;[32] they are increasingly embedded in autonomous weapons

[27] Cathy O'Neil, *Weapons of Math Destruction* (Penguin, 2016); Solon Barocas, Moritz Hardt, Arvind Narayanan, *Fairness and Machine Learning: Limitations and Opportunities* (MIT Press, 2023); Joy Buolamwini, *Unmasking AI: My Mission to Protect What is Human in a World of Machines* (Penguin Random House, 2023); Ruha Benjamin, *Race after Technology: Abolitionist Tools for the New Jim Code* (Polity, 2019); Virginie Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor* (St Martin's Press, 2018).

[28] Andrew Selbst and Solon Barocas, "The Intuitive Appeal of Explainable Machines" 87(3)(2018),pp.1089-1139; Kate Vredenburgh, "The Right to Explanation." *Journal of Political Philosophy* 30(2)(2022), 209-229

[29] Solon Barocas and Helen Nissenbaum, "Big Data's End Run around Anonymity and Consent." in Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum (eds), *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (Cambridge University Press, 2014), pp.44-75.

[30] E.g. Open AI's ChatGPT, has been estimated to consume the equivalent amount of energy as 33,000 households, and the cooling of processors used by Generative AI systems requires vast quantities of fresh water Kate Crawford, 'Generative AI is guzzling water and energy', *Nature* 626 (2024), p.693.

[31] Martin Ford, *Rise of the Robots: Technology and the Threat of a Jobless Future* (Basic Books, 2015); Tapani Rinta-Kahila et al., "The Vicious Circles of Skill Erosion: A Case Study of Cognitive Automation," *Journal of the Association for Information Systems*, 24(5)(2023), pp. 1378-1412

[32] Shoshana Zuboff, *The Age of Surveillance Capitalism* (Profile Books, 2018); Ermelinda Rodilosso, "Filter Bubbles and the Unfeeling: How AI for Social Media Can Foster Extremism and Polarization". Philosophy & Technology, 37 (2024), pp. 1-21.

systems that threaten to cause devastation to human lives, among many others potential harms.[33]

But, equally, AI systems can help further human flourishing, for example, by assisting us in attaining socially valuable forms of scientific understanding; by facilitating efficiency gains and higher standards of service in domains such as education, health care, and the administration of justice; by taking over tasks that are socially necessary yet tedious, unpleasant, or dangerous, liberating people to engage in more meaningful and enjoyable endeavours; and in various other ways we have barely begun to explore, partly because the development of AI has not been primarily steered by those motivated to advance the common good.

Our aim in this section is not the ambitious one of answering all these difficult ethical challenges, but rather to articulate a key notion the Aristotelian framework supports for addressing them: the idea of AI systems as 'intelligent tools' in the hands of democratic citizens animated by the shared project of pursuing their common good. We begin by contrasting human and AI capabilities, before proceeding to explain the notion of AI systems as intelligent tools. We pursue the implications of that notion for two vital domains for personal and communal flourishing: work and democratic politics.

AI systems can have a place in our lives, both as individuals and communities, as intelligent tools, as instruments for human use, rather than as systematic replacements for human endeavours. Aristotelian ethics shows us the point of AI, what AI is *good for*: a tool for advancing human flourishing, for enabling human beings to employ our capacities more freely and fully. We should not try to create AGI, machines with human (or superhuman) reason, and if we did, we could not morally use them as tools (even if we retained the power to do so). Doing so would be to replicate Aristotle's fundamental errors about "slaves by nature." That great philosopher's failure to justify slavery, on the grounds that some humans were nothing more than intelligent instruments, and therefore could, indeed should, justly be employed as tools by "complete humans," is a datum that anyone concerned with ethics in AI ought to take on board.

[33] Birgitta Dresp-Langley, "The weaponization of artificial intelligence: What the public needs to be aware of" *Frontiers in Artificial Intelligence,* 6 (2023)

Humans and AI systems compared

Making sense of AI systems within an Aristotelian framework requires some basic understanding of the nature of such systems and how they differ from human beings in ethically salient respects. Major challenges here include the fact that Artificial Intelligence is a rather amorphous concept, that developments within this field are both highly diverse in character and occurring at a rapid pace, and that there is considerable dissensus among technical experts as to the potential of existing AI methodologies, in particular, with respect to the feasibility and time-frame of achieving the tech industry's Holy Grail of Artificial General Intelligence. Therefore, in drawing a contrast between human nature and AI systems, we must be careful to avoid the naïve and frequently discredited assumption that *a priori* constraints can be readily identified on the capabilities AI systems might acquire in the future. Instead, we will focus on a comparison with AI systems as we broadly know them today, and as we realistically expect them to evolve in the not-too-distant future.

The field of Artificial Intelligence involves the development of algorithms embodied in computer programs. These algorithms can simulate functions that normally require intelligence when done by humans, such as identifying an image as that of a malignant tumour, translating from one language to another, assessing the risk of a creditor defaulting, writing a poem, and so on. Algorithms are mechanical procedures for solving a given problem by means of a finite series of steps. They are 'mechanical' procedures in the sense that they require no resort to judgement in their operation; every step in the procedure is precisely determined. Moreover, what we now call Artificial Intelligence is a form of technology capable of simulating the relevant human functions in a way that exhibits a considerable degree of autonomy, at least in the minimal sense that the operations of such systems do not require human guidance beyond a certain point and can produce outcomes that are neither fully controlled nor predictable by the designers and deployers of these systems.

What is known as classical, or Good Old Fashioned AI operated with algorithms that could be stated in ordinary, natural language. In so-called expert systems, these algorithms sought to crystallise the knowledge of professionals in domains like law or medicine in a series of mechanically-applicable rules. For example, a rule such as, '*if* a person is under 18 years of age on polling day, *then* they are ineligible to vote'. One important benefit of this approach is that it operates according to rules and chains of reasoning that humans can readily grasp. But despite some success in domains such

as chess and routine business administration, by the late 1980s classical AI as a research programme ran aground. The approach proved excessively formalistic and rigid for domains characterised by ambiguity and unpredictability, such as natural language translation and visual object recognition; indeed, for essentially the great majority of human activities.

By contrast, the dominant techniques within the newly emergent Artificial Intelligence of the last few decades are various forms of Machine Learning. This approach involves creating algorithms by training them to identify statistical patterns in vast quantities of digital data. For example, feeding the algorithm data consisting of millions of images of cats and other animals so that it can learn to recognise cats in new data sets. This data-dependence is why the leading AI companies, like Google and Amazon, are those that control huge amounts of data. In Machine Learning, algorithms are configured so as to optimise for some mathematically specified goal, such as shortest travel distance to a destination, risk of re-offending, or the antibiotic potential of a molecule. Because they identify highly complex statistical patterns that can elude humans, Machine Learning systems can generate novel solutions to problems, even astounding their designers. Recall here the famous case of AlphaGo's move 37 in its second match against the world champion of Go, Lee Sedol, a creative move that has been described as one that no human Go player would ever make.

This new generation AI has yielded undeniably impressive results, with remarkable progress being made in areas such as visual recognition, natural language understanding, content recommendation, medical diagnosis, and scientific research. But the enhanced performance of Machine Learning AI systems comes at various costs, of which the following are only a sample. Their dependence on vast stores of data raises serious questions around privacy and intellectual property rights, while their dependence on immense amounts of energy for training and operating models and water for cooling processors used by AI systems has a significant environmental footprint.[34] The immense cost of training AI systems, in terms of data and compute power, raises questions about whether such resources might be better deployed in alternative ways as well as legitimate fears about the risks attendant upon the concentration of such greater power in the hands of a small number of tech companies whose incentives are not obviously aligned with the common good. Moreover, precisely because their operations can outrun humans' understanding, the process through which these systems generate their outputs can be opaque even to their

[34] Kate Crawford, 'Generative AI is guzzling water and energy', *Nature* 626 (2024), p. 693.

designers, the so-called ‘black box’ problem, potentially consigning us to a position of infantile dependency.

It would be foolhardy for anyone, not least humanities scholars, to pronounce *a priori* on the limits to the development of Artificial Intelligence and on whether the goal of Artificial General Intelligence is feasible in the long-term. Still, looking at existing AI systems, and how they are likely to develop in the near future, there are fundamental differences between human intelligence and Artificial Intelligence. For many, at the root of those differences lies the phenomenon of consciousness – a feature (so far) of human, but not artificial, intelligence. But the notion of consciousness is open to vastly different interpretations and, on some minimal specifications of it, such as sentience, it is a quality also shared by non-human animals.[35] On an Aristotelian view, the core contrast is the rational capacities possessed by humans, capacities that may require consciousness as a necessary condition, but are not to be identified with consciousness. And perhaps the key concern an Aristotelian would press in this connection is one succinctly put by the Harvard philosopher Hilary Putnam some decades before the current AI revolution: “The question that won’t go away is *how much what we call intelligence presupposes the rest of human nature*”.[36]

Our intellectual capabilities are rooted in the fact that we live a human life among other humans: we share a world with other humans with whom we also share a biological nature and membership of various communities with their established practices and traditions. Through immersion in this social world we acquire competence in the use of languages with their vast storehouse of different kinds of concepts. Mastery of these concepts enables us to form, justify, and communicate beliefs about how things stand in the world and also to make, and act on, judgments about how things ought to be. We are able to stand back from the promptings of our inclinations, peer pressure or social consensus in order to consider what we ought to believe or do in light of all the reasons, pro and con, that apply to the matter at hand. If challenged about our beliefs or choices, we have the capacity to grasp the meaning of the challenge, to enter into dialogue with our interlocutors, and to explain our beliefs and choices in light of the reasons we believe justify them. We can do this, in part, because we and our interlocutors are engaging with a common world that enables us to appeal to mutually

[35] John Campbell, *Causation in Psychology* (Harvard University Press, 2020).
[36] Hilary Putnam, ‘Much Ado About Not Very Much’, *Daedalus* 117 (1998), 269-281, p.277 (italics in the original). For some other writings which chime with the thoughts developed in the next few paragraphs, see Brian Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (MIT Press, 2019); Erik J. Larson, *The Myth of Artificial Intelligence: Why Computers Can’t Think the Way we Do* (Harvard University Press, 2021); Nigel Shadbolt and Roger Hampson, *As If Human: Ethics and Artificial Intelligence* (Yale University Press, 2024).

intelligible considerations, such as the outcome of an experiment, instructive personal experience, or a compelling narrative of human success or failure. Of course, this is a description of human capabilities, and an ineliminable part of the human condition is the fact that we are fallible, that we can go wrong in exercising these capabilities due to the subversive influence of prejudice, self-interest, inertia, carelessness, cowardice, and various other flaws to which humans are prey. But on an Aristotelian view, we do have these fallible capacities and the genuine possibility of their effective exercise is inherent in our nature and forms the essential backdrop for the enterprise of living a flourishing life in community with others.

Consider, by way of contrast, the nature of 'intelligence' displayed by large language models, such as OpenAI's GPT-4, Google's Gemini, and Meta's LLaMA models. They do not share our biological nature, nor do they have lives to live in a world shared by us. Instead, they are fundamentally highly complex models that generate statistically probable continuations of word sequences based on the distribution of trillions of tokens in the vast corpus of largely human-generated text on which they have been trained. As the computer scientist, Murray Shanahan, has emphasised, attributing mental states such as belief, knowledge, understanding, and communicative intent to such models is problematic even if, with their increasing power and versatility, doing so is a convenient terminological short-hand:

> Humans are the natural home of talk of beliefs and the like, and the behavioural expectations that go hand-in-hand with such talk are grounded in our mutual understanding, which is itself the product of a common evolutionary heritage. When we interact with an AI system based on a large language model, these grounds are absent, an important consideration when deciding whether or not to speak of such a system as if it "really" had beliefs.[37]

The differences at issue here are crystallised by the way in which large language models not only 'hallucinate' – generating false claims that often have an air of plausibility – but more importantly, by the way in which they lack anything approaching what we would regard as common sense. Hence the proneness of AI systems to make spectacular errors no normal human being would ever make.[38] These differences with respect to capacities such as belief-formation and understanding are exacerbated

[37] Murray Shanahan, "Talking About Large Language Models", *Communications of the ACM* 67(2)(2024), p.6

[38] See the TED Talk by a leading expert on AI and common sense, Yejin Choi, aptly entitled 'Why AI is Incredibly Smart and Shockingly Stupid', Available at: https://www.ted.com/talks/yejin_choi_why_ai_is_incredibly_smart_and_shockingly_stupid?subtitle=en (accessed June 14, 2024)

when we consider practical reasoning. Even those thinkers who believe in the imminence of AGI still operate with an essentially means-end conception of intelligence, one that is premised on an AI system achieving a given goal, but not exhibiting the capacity to stand back from any goal and ask whether, in light of ethical reasons, it is a goal it *should* be pursuing or the *moral constraints* on how it should be pursued. Hence the worry that a 'superintelligent' AI may obliterate humanity if doing so is the most efficient means to achieving its goal of increasing the number of paper clips.[39]

From an Aristotelian perspective, our human nature is the basis for a conception of human flourishing as constituted in large part by the cultivation of the virtues, excellences of intellect and character, such as practical wisdom, justice, and courage. AI systems, as we currently know them and as they are liable to evolve in the foreseeable future, lack fundamental aspects of such a human nature, hence also the capability for acquiring virtues constitutive of human flourishing. At best, they are capable of *simulating* virtuous activity in various ways, and will increasingly be able to do so convincingly, a fact that only underlines the importance of retaining a vivid sense of their categorical difference from human beings and their status as tools. They might be able, in various ways, to do what a virtuous person would do, but from the settled disposition to act *for the sake of* the reasons that make it the right thing. As Nigel Shadbolt and Roger Hampson have written recently about the dangers of anthropomorphising AI systems as 'virtuous' agents:

> [I]n principle any virtue, expressed in language or explicit decisions, can be simulated. Although at present these simulations are only partly convincing, that will change reasonably soon. AIs today can also effectively be deputies for us in situations that call for those virtues. What will not change in the foreseeable future is an AI's inability to truly possess a virtue. A soldier who defuses a bomb has courage. The robots already deployed by security forces around the world to defuse bombs do not have courage. The difference is that robots do not feel pain. Nor fear death. Nor experience shame, guilt, or social opprobrium. They don't have agency, autonomy, or sentience. The better they are at embodying human virtues, the greater the accuracy (and sometimes the value) of the

[39] Stuart Russell, *Human Compatible: AI and the Problem of Control* (Penguin, 2020), p.167. See discussion of the 'paperclip maximiser' in Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014), p. 123-124.

> simulation, the more important it will be to define the ways in which they are not the real thing.[40]

Only by grasping the fact that AI systems differ from humans in fundamental ways, thereby warding off the temptations of anthropomorphism, will we be able to make best use of these systems as 'intelligent tools' that facilitate, rather than impede and undermine, our flourishing.

To reiterate, nothing we have said above is meant to suggest that we rule out the possibility of future AI systems developing capabilities for understanding or practical reasoning truly comparable to those possessed by humans. Perhaps this could be done by brute force on the basis of existing methodologies given ever-increasing amounts of data, parameters, and compute power; perhaps this could be done by integrating large language models into wider architectures that provide a semblance of human experience of the world, e.g. robots that physically engage with the world; or perhaps some totally new AI methodology will emerge, distinct from existing machine learning methods that holds out greater promise.[41] But even if something along these lines cannot be ruled out, it is not where we are now. Moreover, it is unlikely that the kind of AGI contemplated here would have anything recognisable as a human nature, with its characteristic profile of capabilities and limitations. This raises deep and unfathomable questions about what the good of such beings would consist in, and what kind of morality would be appropriately tailored to a being of that kind (recall here our reference to the alien morality of the ancient Greek gods in Part I). And this leads to our next concern – the perils of seeking to create such beings in the first place, a matter on which the tragic fallacies and confusions in Aristotle's discussion of "natural slaves" offer powerful grounds for caution.

[40] Nigel Shadbolt and Roger Hampson, *As If Human: Ethics and Artificial Intelligence* (Yale University Press, 2024), p.96.

[41] For the claim that we need to go beyond existing data-driven and statistical methodologies so as to encompass human qualities such as common sense, see Gary Marcus and Ernest David, *Rebooting AI: Building Artificial Intelligence We Can Trust* (Vintage Books, 2019).

Aristotle's failed theory of "natural" slavery

In *Politics* book 1, Aristotle struggles to articulate a theory of "natural" slavery, predicated on the existence of a category of human who possesses intelligence, but is incapable of recognising intrinsically valuable ends, and equally incapable of articulating the right reasons for choosing among available means to a given end: In brief, the "natural slave" is incapable of basic kinds of *moral* reasoning, and thus incapable of human flourishing.[42]

Given his putative incapacities, the natural slave requires direction from a "complete" human master. The relevant context Aristotle says, is a higher order instrument that employs lower order instruments:

> every assistant is, as it were, a tool that serves for several tools; for if every tool could *bring to completion its own work when ordered*, or *by perceiving what to do in advance*, …if thus shuttles wove and plektra played lyres of themselves, then master-craftsmen would have no need of assistants and masters no need of slaves. (*Politics* 1253b33-54a1)

In the absence of these intelligent mechanical contrivances, the slave is necessary as an animate tool that employs subsidiary tools to do necessary work - a tool that, in a puzzling turn of phrase, "shares in reason (*logos*) so far as to perceive it but not to possess it" (*Politics* 1254b22-23).

The central puzzle concerns cognition – what kind of intelligence does an ensouled instrument possess? Aristotle claims (*Politics* 1260a11) that slaves lack the capacity for deliberation, and so cannot grasp the right reasons for action. But they can, he thought, through rational admonition, be taught to employ their intelligence to carry out a complex sequence of actions that promote a goal set by a master, along a path that has been chosen, for the right reasons, by the master. The tensions within Aristotle's account are evident in his terminology of "dead ends," "shocking conclusions" and "peculiarity" in reference to issues that his theory of natural slavery raises. And his elaborate ethical-cognitive edifice collapses of its own weight when put into practice: Strikingly, Aristotle admits that observers cannot distinguish between a natural slave and a complete human. The slave cannot be identified by physical traits; the slave uses language, feels pleasures and pains, possesses certain (albeit delimited) virtues, and is

[42] See, further, Josiah Ober, "Ethics in AI with Aristotle," paper presented at the Oxford Centre for Ethics in AI, June 16, 2022.

capable of complex sequences of (albeit non-moral) means-ends reasoning. As such the slave is identical, in many salient ways, to many other humans.

Why then did Aristotle go to such length to defend a theory of natural slavery? A theory by which – contrary to his own experience– the interests of slaves and masters were congruent, and their relationships friendly and just. The answer is that Aristotle saw no hope of creating a just social order, one which provided the external goods necessary for flourishing, *without* ensouled instruments. Since he believed that working under another's supervision impeded the development of virtue, yet such labour was required for the preservation of social welfare, he theorised instruments that benefitted by servitude. He insisted, against all he knew about the minds of actual enslaved persons, and the actual relations between slaves and masters, that ideal-type instruments, with appropriately defective forms of reason, existed in the phenomenal world. Without such instruments, his ideal community must remain a fantastic utopia, so, he willed them into being, in the face of all the contradictory evidence, and to the detriment of his own political theory.

Aristotle's failed theory of natural slavery is instructive for the ethics of AI both because it elucidates problems that arise with human-like AGI and because it introduces the concept of the 'intelligent tool' properly used to promote human flourishing, a tool that comes into play only after moral ends and means have been set. Humans use moral understanding to determine valuable ends and to set the frame for how those ends are to be pursued. Ends must be pursued deliberatively: "in the right way, by the right actions, at the right time." We choose the right ends and means to those ends through prosocial, virtuous employment of our capacities of reason and communication. AI, as a tool, ought to be used to enhance our ability to act effectively, in prosocial, virtuous ways: to help us to develop and manifest our virtues of courage, wisdom, moderation, justice; and also generosity, piety, mercy, etc.

AI might, for example, aid human creators in developing and expressing new forms of visual art, performance, writing etc. AI is not inherently creative, but could be developed as a tool for enhancing human creativity – just as other artist's tools (e.g. technology for casting bronze) have enabled the expression of new forms of art. In what follows, we focus on the idea of AI systems as tools in two domains of human life: work and democratic self-government.

In the realm of work, we are free to reject Aristotle's belief that working at the direction of others, or for the sake of others' enjoyment (e.g. in musical performance: *Politics* 1341b9-14), is inherently illiberal, and as such degrades our moral capacity. Work is a profoundly important context for human fulfilment. Engaging in work activities can have instrumental value, for example, by generating valuable goods and services and an income for the worker or by honing skills that are useful beyond the workplace, such as in personal life and political activity. But it can also be a source of great non-instrumental value that is constitutive of human flourishing. Key among these values is that of achievement, which consists in the valuable exercise of our powers in meeting difficult challenges for a worthwhile end. Meanwhile, collaborating with one's work colleagues and customers towards the realisation of a common project is also an important form of what Aristotle called civic friendship (*NE* VIII.9), which includes both mutual benefit and genuine concern for the good of one's partners in an important enterprise. A politically significant by-product of achievement is a justified sense of self-esteem, a quality that can play a vital role in sustaining a democratic ethos in which citizens feel able to look each other in the eye as equals, contributing through work to the common good of society as a whole. This depends on a general recognition of the valuable role that different forms of work, whether of hand, brain, or heart, play in sustaining the common good.

But we can nonetheless acknowledge that much work has, historically, been degrading both physically and psychically – dull, repetitive, unimaginative, exhausting. AI can help us to eliminate degrading labour by transforming work into an expression of our prosocial capacities, our human excellences. As with other kinds of activity, work ought to be for the ultimate end of joint and several flourishing. The products of labour ought to promote rather than impede flourishing.

If work is a vital domain for human flourishing in the modern world, the question arises: How should we respond as AI systems increasingly acquire the capacity to perform work activities? Of course, there are many forms of work – notably, dangerous or demeaning or otherwise distasteful tasks – that we should be happy, perhaps even obligated, to delegate to AI systems. But there has never previously been a technology that has the potential to replace human work activities on such a significant scale. This goes well-beyond 'routine' or 'mechanical tasks' to include white-collar occupations, such as journalism and legal services. A 2017 Oxford-based study, conducted *before*

recent breakthroughs in AI, concluded that 47% of all occupations in the United States are capable of being "computerised" in the next 10-20 years.[43] One needs, however, to take such dramatic claims with a grain of salt. As Oren Cass has noted, one of the jobs the authors of this study supposed could be fully automated is that of school bus driver:

> From a tall enough ivory tower, or a heady corner of Silicon Valley, the claim about school bus drivers might seem to make sense. What could be easier than driving a school bus? The route is the same every day, it's short, and it gets cancelled for snow. For parents, though, the idea of locking twenty kids in a self-driving vehicle for half an hour, with no adult supervision, sounds dubious at best.[44]

More circumspectly, a McKinsey study in 2018 found that, although just under 5% of occupations are fully automatable, around 30% of all work tasks in 60% of occupations could be automated.[45] Given dramatic advances in LLM's since these studies, the potential threat posed by AI to jobs seems significant. Nor can we confidently assume that, as with past technological innovations, new jobs will emerge to replace those eliminated by AI.

Among tech leaders, such as Mark Zuckerberg and Elon Musk, the idea of a Universal Basic Income has found favour as a solution to the challenges of a post-work world.[46] Now, there is reason to doubt that the replacement of human workers with AI systems will lead to productivity gains large enough to generate a UBI that covers lost income – as we have seen, a great deal of automation represents what Acemoglu and Johnson have called 'so-so innovation', displacing workers without any significant rise in productivity or the quality of goods and services. Much here, of course, turns on the contested issue of whether Artificial General Intelligence is a feasible, desirable, and reasonably imminent development.[47]

---

[43] CB Frey and MA Osborne, 'The Future of Employment: How Susceptible Are Jobs to Automation?', *Technological Forecasting and Social Change* 114 (2017): pp. 254-80.

[44] Oren Cass, *The Once and Future Worker: A Vision for the Renewal of Work in America* (Encounter Books, 2018), p,69.

[45] James Manyika and Kevin Sneader, "AI, Automation, and the Future of Work: Ten Things to Solve For," McKinsey Global Institute Executive Briefing, June 1, 2018, https: //www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-fo.

[46] Catherine Clifford, 'Elon Musk: Robots Will be Able to Do Everything Better Than Us', CNBC, July 17, 2017 https://www.cnbc.com/2017/07/17/elon-musk-robots-will-be-able-to-do-everything-better-than-us.html

[47] In a recent paper, Daron Acemoglu has argued that the GDP gains generated by AI will be in the range of 0.93%-1.16% over ten years in total. Moreover, he contends that greater gains will require a 'fundamental reorientation' of the AI industry away from 'general human-like conversational tools' to tools that can increase the marginal productivity of workers by providing them with reliable information. See Daron Acemoglu, 'The Simple Macroeconomics of AI', *National Bureau of Economic Research* Working Paper 32487 (May 2024), p.45.

But even if a UBI could address the economic inequalities threatened by technological unemployment by redistributing shares of a larger economic pie, it would not itself remedy the massive reduction in opportunities for achievement, as well as the consequent loss of a source of friendship and democratic self-esteem. And, from a purely practical standpoint, it seems unlikely that the UBI scheme would be sustainable in the long run. This is because we could imagine a democratic ethos being frayed by the status inequalities and the attrition of skills consequent upon the displacement of large numbers of citizens from productive activities.

One obvious thought here is that the loss of opportunities for achievement will be compensated for by the greater leisure time people will have to pursue other values in a world without work, values such as personal friendship, the pleasures of artistic appreciation, travel or fine dining, religious observance, etc. John Maynard Keynes greeted the prospect of a jobless future in this vein in his 1930 essay entitled 'Economic Possibilities for our Grandchildren' – claiming that "we have been trained too long to strive and not to enjoy".[48] But even this cautious optimism seems overly optimistic to us, elevating passive enjoyment to a status it cannot have in an Aristotelian framework that prioritises the cultivation of the excellences of character and intellect in its account of human flourishing.

Another alluring thought advanced recently is that the lost opportunities for achievement afforded by work can be replaced by achievement in the playing of games.[49] Some have taken this idea even further, suggesting that we will be occupied in playing even more exciting and challenging games in virtual reality, involving such things as the virtual capacity to fly unaided, and that a life spent doing so can be just as good as a life spent in the 'real world'.[50] We have here, it seems, the elements of a philosophical prescription for the good life in the metaverse.

But this play-based virtual utopia rests on deeply contestable theses. First, that the primary value of game-playing is that of achievement, rather than play itself (when factory workers play an impromptu game of football in their lunch-break, is the primary

[48] John Maynard Keynes, 'Economic Possibilities for our Grandchildren', in *Essays in Persuasion* (Harcourt Brace, 1932), pp. 358-373

[49] See John Danaher, in *Automation and Utopia: Human Flourishing in a World without Work* (Harvard University Press, 2019), p.236, referencing the work of Thomas Hurka, 'Games and the Good'. *Proceedings of the Aristotelian Society* supp. Vol.80 (2006), pp.217-35.

[50] For the philosopher David Chalmers, "virtual realities have comparable value to nonvirtual realities", hence life in a computer simulation can be just as meaningful as life in a non-virtual world. D. Chalmers, *Reality+ Virtual Worlds and the Problems of Philosophy* (Allen Lane, 2022), p.328

value realistically secured thereby that of achievement in athletic prowess, rather than the fun of a kick-about?). Second, that the value of achievement is to be understood in an explicitly anti-Aristotelian way, as focussed entirely on the process involved (skill in overcoming difficulties) rather than also on any substantive good achieved by that process. After all, the typical objectives of games (e.g. putting a ball in a hole, crossing an arbitrary line before others) may be trivial or valueless. Only when the value of ends comes into view – the valuable goods and services produced through work – can we grasp why the sense of achievement derived through work as a nurse, plumber, teacher, farmer etc. cannot for the great majority of people be satisfactorily replaced by proficiency at activities such as chess, golf or table-tennis.[51] And it is precisely the disconnect from suitably valuable ends that explains why many fail to find fulfilment in otherwise well-paid, but pointless, white collar 'bullshit jobs', however challenging or difficult the tasks they involve.[52]

Finally, the thesis about virtual reality brings us back to even more fundamental themes broached in Part I, about our essential nature as human beings, with a specific biological form, one adapted to living in the natural world along with other human and non-human beings with their own biological form, and of work itself not just as a source of achievement, friendship and self-esteem, but also as a way of engaging with and understanding a reality that exists independently of us rather than an artificial reality synthesised by human beings. As David Wiggins has written, elaborating a neo-Aristotelian account of the meaning of work:

> Acts or activities that apply what he [Aristotle] calls a rational principle aim at something worthwhile by drawing upon faculties and dispositions whose exercise gives pleasure (a distinctive, associated pleasure) to the doer and enlarges also – here I reach beyond Aristotle – the doer's understanding of the realities we inhabit. That is to say that the exercise of these faculties or dispositions affords both practical understanding of those realities and the satisfactions that we attain by learning to wrestle or struggle with them.[53]

The idea here is that work affords a distinct form of understanding that emerges through contact with an independent and potentially recalcitrant physical reality: in

[51] For these two objections, see John Tasioulas, 'Games and the Good', *Proceedings of the Aristotelian Society* supp. Vol. 80 (2006) 237. In line with the first objection, however, play is a substantive good itself that the playing of games can realise. See also John Tasioulas, 'Work and Play in the Shadow of AI', in David Edmonds (ed), *AI Morality* (OUP, forthcoming).
[52] David Graeber, *Bullshit Jobs: A Theory* (Allen Lane, 2018).
[53] David Wiggins, 'Work, its moral meaning and import', *Philosophy* 89 (2014), p.479

work, we engage with that reality in a way that deploys our rational faculties in order to create goods and services that satisfy human interests. This is related to the idea found in Hegel's dialectic of master and slave, which in turn influenced Marx, that work is vitally significant for human self-realisation because it involves the struggle to transform nature into a humanised domain of culture (Bildung).[54] By humanising nature in this way we more fully realise our own human potential, which includes enabling us to conceive of ourselves as moral agents possessing equal rights.

In short, how AI technology should be integrated as tools into the work environment so as to enable both the flourishing of workers and the common good of society remains a serious challenge. The simple idea that we should promote the use of AI technology, including by systematically displacing human workers, in order to stimulate economic growth and then redistribute some of the wealth gained to the jobless through a UBI, turns out to be dangerously simplistic. A focus on economic growth, through such measures as GDP, fails to reflect all the value we derive from work, notably, achievement, friendship, self-esteem, and an understanding of the world around us by engaging with it to produce goods and services.

Reverting to Aristotle's scepticism about a life of work due to its supposed lack of self-direction, an important element in addressing this challenge will be giving workers a greater voice in determining the shape of their workplace. This includes the role that AI technologies should play at work. One of those beneficial roles for AI technology should be that of enhancing worker participation in corporate governance, as opposed to its being part of a system of surveillance and control geared to extracting the maximum economic value from 'human resources'. Consider, for example, the German co-determination system, whereby workers' councils have a say on a company's supervisory board. The means of production ought to freely enable us to express our capacities, meaning that workplaces ought to be democratised. Workplace democracy is potentially highly productive.[55] AI ought to be used to enable workers to contribute to work processes and products using their heads as well as their hands - inverting Henry Ford's (perhaps apocryphal) regret that the hands he hired in his automotive factories came attached to heads. AI tools should enable workers to be participants in the organisation and management of work environments, in much the same ways that it could enable more participatory forms of political democracy. By the same token, AI that threatens to degrade the free expression of human capacities, which makes

[54] G.W.F. Hegel, *The Phenomenology of Spirit* (Oxford University Press, 1976), pp.111-119.

[55] Detailed, for the pre-AI environment, in Brook Manville and Josiah Ober, *A Company of Citizens* (Harvard Business School Press, 2003).

workplaces more hierarchically centralised and less democratic, deserves to be regulated.

Participatory democracy

As we have seen, Aristotle's argument from fundamental human nature leads to the conclusion that participatory democracy is the form of government most conducive to human flourishing. But can a participatory democratic process produce good policy that supports the material well-being that is an essential foundation for moral flourishing? And how might AI help or hinder the goal of scaling up participation to the level of the nation state, or beyond?

There is substantial evidence that, in modernity, democracy is positively correlated with, and a driver of, increased state capacity, economic development, and higher levels of consumption.[56] Consider, further, in this connection, empirical findings to the effect that there has never been a famine in a modern democracy,[57] that democratic government is the most important factor in upholding human rights,[58] and that democracies tend not to go to war against each other.[59] In Greek antiquity, at least in the uniquely well-documented case of ancient Athens, there is reason to believe that relatively robust economic growth, accompanied by relatively low income inequality, was promoted by distinctively participatory democratic processes.[60] Aristotle appears to be aware of that potential. One important line of argument for democracy, developed by Aristotle in a famous passage (*Politics* 3.11) is explicitly outcome-driven: Under the right conditions (assuming a well educated, so adequately virtuous citizenry), a large and epistemically diverse body of persons is more capable of making objectively correct decisions on important matters relevant to community well-being than any small body of experts. It has that capability because, on the analogy of a pot-luck dinner that is superior to a meal provided by a single benefactor, the diverse group can draw on different forms of useful expertise, in the form of experience, information, and knowledge.[61] That the citizens themselves understood the effectiveness of democratic processes is implied by Aristotle's comment in the *Politics*,

[56] Daron Acemoglu, Suresh Naidu, Pascual Restrepo, and James A. Robinson, 'Democracy Does Cause Growth', *Journal of Political Economy*, 127 (2019), pp. 47-100.
[57] Amartya Sen, *Poverty and Famines: An Essay on Entitlement and Deprivation* (Oxford University Press, 1983).
[58] Kathryn Sikkink, *Evidence for Hope: Making Human Rights Work in the 21st Century* (Princeton University Press, 2017).
[59] Michael Doyle, 'Kant, Liberal Legacies, and Foreign Affairs: Parts I and II', *Philosophy and Public Affairs* (1983), pp. 323-353.
[60] Josiah Ober, *Democracy and Knowledge* (Princeton University Press, 2008)
[61] Josiah Ober, "Democracy's Wisdom," *American Political Science Review* 107(1) (2013), pp. 104-122.

that sometimes the person who lives in a house knows more about it than the architect who designed it.[62]

Aristotle himself never participated as a citizen in a democratic community. But he chose to live for most of his life under democratic Athenian law, as a resident foreigner in Athens.[63] As the Aristotelian text, *The Constitution of the Athenians* (probably written under Aristotle's direction, by a student at the Lyceum), and passages in the *Politics* and *Nicomachean Ethics* attest, he was a careful, if critical, observer of the political institutions through which the Athenian citizens (some 30,000 adult men in Aristotle's time) "ruled and were ruled over in turn." As with Aristotle's ethics, Athenian democratic institutions, which (*inter alia)* excluded women and employed public slaves, are clearly unsuitable for wholesale modern adoption. But once objectionable features have been excised, Athenian institutions offer insights into how AI could promote human flourishing, by engaging citizens in meaningful activity of self-government.

At the level of fundamental principles and background conditions, Athenian democracy, as Aristotle clearly recognised, exemplified the values of political equality, free speech, and civic dignity. Moreover, Athenian public culture was, in many ways, an education in democratic citizenship: from childhood through the course of his life, the Athenian citizen learned how to perform the duties of a citizen, and was given reasons to believe that the costs of those duties were well worth paying.[64] That education began with the family and the local community: the Athenian approach to collective self-governance employed the principle of subsidiarity, in that decisions with specifically local impact were made locally, at the level of the "demes" – the 139 towns, villages or neighbourhoods of Attica. Deme assemblies confirmed the citizen status of male residents upon their coming of age, and occasionally considered challenges to citizen status.

At the state level, Athenian policy-making began in a deliberative Council of 500 citizens over age 30, chosen by lottery for a one-year term. The selection process ensured that men from all 139 demes served every year as council-members; because lifetime Council service was limited to two non-consecutive terms, a high percentage of

[62] *Politics* 1282a14-22.

[63] A longtime resident resident foreigner in; Athens, Aristotle lived for a time with autocrats, King Philip II of Macedon (as tutor to his son, Alexander III - "The Great"), and with his father-in-law, Hermias, the semi-independent ruler of Atarneus. Aristotle left his birth-polis of Stagira before coming of age, and never returned; Stagira was destroyed (and perhaps refounded) by Philip II, during Aristotle's lifetime.

[64] Josiah Ober, 'The Debate Over Civic Education in Classical Athens', in Yun Lee Too, ed., *Education in Greek and Roman Antiquity* (Brill, 2001), pp. 273-305.

all Athenian adult men served (including, for example, the philosopher Socrates). The annual membership of the Council consisted of 10 "tribal" teams of 50. Each was composed of 17 or 18 men from inland, coastal, and urban demes, thereby ensuring the geographic diversity of every tribal team. Each tribal team, and the Council overall, thus approximated a representative sample of Athens' citizen population. Each of the 10 teams presided, in rotation, over plenary Council meetings, which in turn set the agenda for 40 annual meetings of the citizen Assembly. Some 6000-8000 citizens attended each meeting of the Assembly; they gathered at dawn in an open-air, purpose-built theatre-like structure. Assembly meetings were presided over by a rotating team of nine Councilmen.[65]

The presiding Council members presented each agenda item (through a herald) along with the Council's recommendation (if any). They then invited "any Athenian who so desired" to speak in favour of the motion, to oppose it, or to amend it. Given time constraints, few citizens addressed the Assembly on a given measure, but many others participated actively by vocally expressing their approval of a speaker – or their disapproval, promptly seeing off those deemed inexpert on the topic at hand. The presiding team sought to identify a proposal that could gain wide support, by attending carefully to the audience's responses to rival proposals and amendments. Most citizens in attendance were experienced judges of public rhetoric, attentive to the distribution of relevant expertise and trusted opinion among both speakers and others in the audience. While citizens had different personal interests, the direct impact of public decisions on community welfare encouraged a primary focus on the common good. The final vote on a proposal was ordinarily by open show of hands. The decision was frequently unanimous, or close to it: The public debate/vocal response/hand voting procedure encouraged 'virtuous cascades' in favour of options widely regarded as the best available. Although bad decisions were sometimes made (and called out as such by ancient critics of democracy), the political process produced policy that overall enabled the Athenian community to prosper over time and to survive periods of crisis.[66]

Moreover, and to the point of the role of democracy in human flourishing, the Athenian process exemplified and reified the values of freedom, equality, and civic dignity: Liberty in the free speech and freedom of association that were the hallmarks of public debate. Equality in the equal vote of each citizen, and each citizen's equal opportunity to be chosen in a lottery, as a member of the Council, as a magistrate, or a member of

[65] P. J.Rhodes, *The Athenian Boule* (Oxford University Press, 1985).
[66] Mirko Canevaro, 'Majority Rule Vs. Consensus: The Practice of Democratic Deliberation in the Greek Poleis', in Mirko Canevaro et al. eds., *Ancient Greek History and Contemporary Social Science* (Edinburgh University Press, 2018), pp. 101-156.

a jury. Dignity in the high standing according to each citizen, his immunity from the dignitary harms of humiliation and infantilization, his expectation of being treated with respect and recognition by his fellow citizens, just as long as he accorded them the same.[67]

The Athenian ideal of participatory democracy may seem unfeasible in the context of enormous and heterogenous modern states that exist in the world today. This supposed fact has often been invoked as a justification for representative democracy, which relies on a class of professional politicians to act as representatives of democratic publics in formal law-making. This mode of democracy, rather than the more radically participatory Athenian variety, is the standard model of among leading political philosophers such as John Rawls and Jurgen Habermas. But contemporary representative democracy today is almost everywhere in crisis; people are increasingly losing faith in it, especially the young in long-established democracies. The crisis seems to be the result, among other factors, of two mutually reinforcing tendencies: technocracy and exclusionary populism. Both of these tendencies are united by their anti-democratic character: rule by experts (judges, central bankers, bureaucrats etc), on the one hand, and rule by an authoritarian strongman embodying the will of the 'real people', on the other. Technocracy, by removing certain matters from democratic decision, sparks an authoritarian populist backlash that also seeks to bypass democratic procedures, which in turn leads to ever more insistent calls for technocratic rule, in an ever-deepening anti-democratic spiral.

To break out of this impasse, we need to re-invigorate a more participatory style of democracy, closer to the Athenian ideal, drawing on technological and other means now at our disposal for scaling up citizen participation in political deliberation and decision-making. Here, AI systems can play a valuable role as we re-imagine what participatory democracy can mean in the twenty-first century. This hopeful prospect needs to be set against the more common image of AI tools as subverting democracy through the spread of misinformation, disinformation, the fomenting of political polarisation, and the enablement of various forms of corporate and governmental surveillance and control.

How might AI enable more participatory forms of democracy today? One way forward is by adapting Athenian political principles and institutions to our contemporary conditions. First, by facilitating subsidiarity: AI could sort salient issues according to

[67] Josiah Ober, 'Democracy's dignity', *American Political Science Review*, 106.4 (2012), pp. 827-846.

the locus of their greatest impact and provide for moderated deliberation among and decision by those most directly affected – whether the relevant locus was geographic or otherwise. Next, in cases where the locus of impact was larger, and the population of those directly impacted too great for efficient “face to face” deliberation and decision, AI could help to select an agenda-setting and advisory council by efficiently identifying truly random samples of larger populations. The chosen members of the council could then be provided with AI generated expert opinion tailored to the issues at hand, enabling them to deliberate among themselves, and make a collective decision accordingly. The basic approach is modelled on a deliberative “mini-public” which sets its own agenda and makes policy recommendations to a larger assembly in which all affected citizens are able to participate.[68]

At the level of the affected-citizen assembly, properly monitored (by responsible human agents) AI could again serve as a 'trusted expert' that would provide essential informational inputs to each citizen, tailored to each individual’s learning processes. Each citizen would receive the same curated information, but in language and format that would be most accessible to each. AI could identify and circulate alternative proposals and measure the depth (intensity) and breadth (numbers) of audience responses to each, driving towards a decisive vote on a measure likely to gain wide support (or be soundly defeated). This approach would potentially enable mass participative democracy, with an informed (civically educated, but not necessarily technically adept) citizenry. Recent experiments with democratic corporate governance, and in Taiwan have demonstrated the potential value of digital tools for participatory democracy; that value could be increased as AI becomes a more powerful tool for advancing human purposes.[69]

AI-assisted democracy, like the Athenian political system, will require certain background conditions: It depends on a citizenry that is willing and able to pay the costs of its own ongoing civic education, citizens who become experienced in collective decision-making, and who are capable of distinguishing common from factional interests. AI could help foster that education, but it cannot create the motivation for it. The AI-generated information must be treated by the citizens as expert assistance, rather than merely echoing and reinforcing prejudices.

---

[68] Hèléne Landemore, *Open Democracy: Reinventing Popular Rule for the Twenty-First Century* (Princeton, 2020).

[69] Taiwan as a model of the use of AI/digital technology to enhance citizen participation: /https://www.radicalxchange.org/media/papers/Taiwan_Grassroots_Digital_Democracy_That_Works_V1_DIGITAL_.pdf, The use of AI to enable more democratic corporate governance https://iai.tv/articles/we-need-to-democratize-ai-helene-landemore-john-tasioulas-auid-2680

In sum, done right, AI could enhance democracy by improving both private deliberations (providing better, more accessible information) and deliberative communication with others (expressing our ideas more clearly and powerfully – i.e. improved rhetoric. It could promote fair and efficient collective decision-making, leading to choices and actions that improve the community – rendering it more prosperous (to a point), more just. AI ought to allow us to do all this at increasing scales at lower cost: with fewer painful sacrifices. In brief, it ought to make democracy more participatory, meaningful, and effective. AI enhanced politics could fulfil the promise of democracy that was at once deliberative, participatory, decisive, and effective; employing expertise for decision-making while avoiding the domination of experts. Done badly, however, AI could enable elite capture by becoming the expert that makes important decisions, or by being monopolised as a tool of an elite. Alternatively, by corrupting political communication, AI could exacerbate polarisation, discord and destructive forms of populism.[70]

[70] For work examining threats to democracy by AI (focusing on communication), see Sarah Kreps, and Doug Kriner, 'How AI Threatens Democracy', *Journal of Democracy*, 34 (2023), pp. 122-131. For the potential benefits, see Hélène Landemore, 'Can AI bring deliberative democracy to the masses?' presented at HAI Weekly Seminar and NYU Law School (2022). Finally, for work on the necessity of civic education for the survival of democracy see Manville and Ober, *Civic Bargain*.

PART III

SIGN-POSTS FOR REGULATION

> [W]hile the state came about as a means of securing life itself, it continues in being to secure the good life.
>
> Aristotle, *Politics* (1256b27-31).

Regulation is the activity of establishing and implementing rules, principles, and policies that enable us to realise our values – ranging from safety and human rights to environmental protection - more efficiently and effectively than would otherwise be the case. These regulations can take a multiplicity of forms, including laws at the domestic, regional, or international level, not all of which may be backed up by effective enforcement mechanisms; 'soft law' norms like the UN Guiding Principles on Business and Human Rights that are not legally binding; industry codes of conduct; social conventions; and personal policies adopted by individuals or associations. One of the merits of a well-constructed regulatory scheme is that it can significantly reduce the need for deliberation prior to action, since adherence to its prescriptions will generally ensure adequate conformity with the underlying values the scheme is intended to promote. Consider, for example, the way in which automatic compliance with the rule of the road – 'Drive on the left' – secures the underlying values of safety and freedom of movement.

One stubborn misconception is that the AI revolution demands a corresponding revolution in our regulatory thinking - that we have to "start from scratch", in the words of the U.S. Senate Majority Leader, Charles Schumer, in regulating AI.[71] But just as the idea that we are faced with an ethical blank slate in dealing with AI is a profound error, if our claims about the enduring significance of the Aristotelian framework are correct, so too is the equivalent claim about regulation. We should not allow the dizzying progress in AI technology to stampede us into assuming that existing regulations are silent on the predicaments posed by AI, even if this attitude might understandably find favour among tech companies who stand to benefit from the freedom it affords them and who would also likely wield an outsized influence in framing any new, bespoke

[71] Quoted in Alondra Nelson, 'The Right Way to Regulate AI: Focus on its Possibilities, Not its Perils', *Foreign Affairs* (Jan 12, 2024). https://www.foreignaffairs.com/united-states/right-way-regulate-artificial-intelligence-alondra-nelson

regulations. Instead, the first port of call must be to consider how existing regulatory frameworks - such as intellectual property regimes, principles governing legal personality and responsibility, administrative law constraints on decision-making by public bodies, human rights norms, environmental law principles, and so on - can be intelligently extended to the challenges posed by AI.[72] In the words of Alondra Nelson, former Director of the Office of Science and Technology Policy in the Biden White House:

> Democratic leaders must understand that disrupting and outpacing the regulatory process is part of the tech industry's business model. Anchoring their policymaking process on fundamental democratic principles would give lawmakers and regulators a consistent benchmark against which to consider the impact of AI systems and focus attention on societal benefits, not just the hype cycle of a new product. If policymakers can congregate around a positive vision for governing AI, they will likely find that many components of regulating the technology can be done by agencies and bodies that already exist.[73]

They may also find that few radical regulatory changes are required, and certainly no discrete body of law worthy of the name 'law of AI', no more than there is a 'law of the horse', for all the tremendous impact of this animal on human history. Rather the emphasis will be on the imaginative application of existing legal principles and regulatory norms in light of the underlying values they are intended to serve.

The note of caution just registered is consistent with acknowledging that the challenges posed by AI technology are potentially so novel - for example, due to the rapid development of radically new capabilities with a multiple-use character - that they will compel the serious re-examination of existing regulatory schemes. This process will require us to delve into the underlying values that such regulation should serve, both the proper specification of those values and their correct prioritisation in different use domains will become live questions. We need to engage with these questions in order to interpret existing standards, such as those pertaining to privacy and intellectual property,[74] in light of the new challenges posed by AI and, more radically, to formulate

---

[72] See, for example, Simon Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law* (Cambridge University Press, 2021).

[73] Alondra Nelson, 'The Right Way to Regulate AI: Focus on its Possibilities, Not its Perils', *Foreign Affairs* (Jan 12, 2024). https://www.foreignaffairs.com/united-states/right-way-regulate-artificial-intelligence-alondra-nelson

[74] See, for example, Effy Vayena and John Tasioulas, 'The Dynamics of Big Data and Human Rights: The Case of Scientific Research', *Philosophical Transactions of the Royal Society* (2016).

new standards in response to some of those challenges, e.g. a right to a human decision.[75]

Moreover, AI is a totalising technology, in the sense that it pervades and has a potentially transformative impact on *all* domains of human life, both public and private. Partly for this reason, the adequate regulation of AI requires that we address its implications for the *full range* of human values. This is why we have approached the ethics of AI by offering a presentation of the core tenets of the Aristotelian framework which is comprehensive in its scope (Part I), and elaborated the idea of AI systems as 'intelligent tools' that emerges from it (Part II). In the rest of Part III we compare the Aristotelian framework as a basis for AI regulation with both the emphasis on 'safety' as an over-arching regulatory goal and the ideological templates that drive the regulatory efforts of the world's three 'digital empires', before proceeding to the issue of the elaboration of global standards for the regulation of AI. In the latter context, we conclude by proposing a novel human right - a human right to a human decision.

Beyond the Rhetoric of Safety

The value pluralism of the Aristotelian is a reason for wariness towards efforts by governments and corporations to subsume AI regulatory endeavours within a single, overarching rubric, such as trust, or the good, or safety. A prominent example is the United Kingdom's international AI Safety Summit at Bletchley Park in November of 2023. The main problem with the safety framing is that it characterises the challenge of regulating AI in unduly narrow terms, as a matter of avoiding threats to certain narrowly circumscribed values (primarily, life and limb). Of course safety so conceived is vitally important, but treating it as the exclusive or primary focus of AI regulation distracts us from the fundamental question of what good AI systems might help us achieve. Often, of course, it is uncritically assumed that the good in prospect is economic growth, an objective whose severe limitations we have already discussed. From an Aristotelian perspective there is no reason to develop and deploy AI systems unless some significant human goods are achieved by doing so, and it is impossible to assess the significance of the risks associated with AI without balancing them against these benefits. In the words of Verity Harding:

[75] Yuval Shany, 'The Case for a New Right to a Human Decision Under International Human Rights Law', Nov. 2023 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4592244 and John Tasioulas, 'A Human Right to a Human Decision' (forthcoming).

> [A]nyone intent on building an AI-powered future should begin with an exciting vision - rooted in tangible public benefit and the values of human rights and democracy - showing what they want to deliver and why it will be great for the world at large. If you can't do that because you don't know what that future should be, then maybe you shouldn't be pushing relentlessly toward it.[76]

Moreover, the safety framing tends to sideline a wider category of risks that do not readily fall within its compass, including more 'intangible' harms, such as violations of rights to privacy, non-discrimination, free speech, and dignified conditions of work.

But there is another problem with all-encompassing rubrics like 'safety', quite apart from obscuring the diversity of values to which AI regulation should be responsive. These overarching rubrics – and 'safety' is a prime example – threaten to present the ethically and politically contentious issues around AI regulation as primarily technical issues to be resolved by experts. It's as if - as with product safety regulations - the objectives that regulation should serve are uncontentious and that the real work that is needed involves deploying scientific and bureaucratic expertise to secure those objectives. What threatens to get lost in this technocratic picture of regulation are two important truths. First, regulation is not purely or even primarily a matter of technical expertise. It involves a whole series of potentially difficult value judgments – identifying the values in play, interpreting their demands in the present context, and striking a sound balance among them in cases where they are in tension, including trading off safety against gains in efficiency or fairness. And, second, contrary to epistocrats from Plato onwards, there is no class of experts when it comes to these ethical questions. Instead, robust democratic processes that involve deliberation and decision-making by an informed and engaged public are needed to underwrite the legitimacy for our regulatory schemes. Experts provide vital assistance to these democratic processes, but they should not displace or commandeer them.

Yet, all too often, the debates around AI regulation are conducted as a dialogue among a narrow set of technocratic elites, with the perspective of ordinary people whose lives are increasingly affected by these technologies consigned to the margins. The deficit of

[76] Verity Harding, *AI Needs You: How We Can Change AI's Future and Save Our Own* (Princeton University Press, 2024), p.223. See also Alondra Nelson, 'The Right Way to Regulate AI: Focus on its Possibilities, Not its Perils', *Foreign Affairs* (Jan 12, 2024) who notes that the issue of positive benefit tends to be overwhelmed by anxieties about catastrophic risks: "But when tackling AI governance, it is crucial for leaders to consider not only what specific threats they fear from AI but what type of society they want to build. The public debate over AI has already shown how frenzied speculation about catastrophic risks can overpower people's ability to imagine AI's potential benefits". https://www.foreignaffairs.com/united-states/right-way-regulate-artificial-intelligence-alondra-nelson.

wider public engagement was epitomised by the heavy presence of representatives from big tech corporations at the Bletchley summit alongside governmental leaders. Indeed, the exclusionary nature of the discussion around AI regulation was perhaps most powerfully symbolised by the interview the United Kingdom Prime Minister, Rishi Sunak, conducted with the tech billionaire Elon Musk, an exchange in which any sharp differences of view (including about work in the age of AI) seemed to be politely shelved.[77]

But it might be objected: what about the fear of existential risk that drives much of the 'safety' discourse in AI circles, isn't the magnitude of this threat something that justifies prioritising safety as the framing for AI regulation? One version of this worry concerns the use of AI by malign actors, such as authoritarian governments or terrorist organisations, and this clearly is a serious threat. But often it takes the form of AI systems themselves autonomously posing an existential threat quite independently of any evil intent on the part of humans involved. To be clear, we do not contend that we should totally discount the worry that AI might spiral out of control and annihilate humanity, either through some eccentric attempt to pursue seemingly innocuous goals we have given it or else in pursuit of its self-given goals. But this scenario not only rests on a series of contestable empirical assumptions, regarding how AI technology is likely to evolve and on what time-scale. It also rests on contestable conceptual assumptions about the very idea of AGI. For example, is it realistic to suppose that true AGI – an AI system that was genuinely capable of replicating human cognitive functioning across its entire range of operation – would be so utterly morally obtuse as to annihilate all human beings in order to fulfil an instruction to produce paper clips?

More importantly, perhaps, the weight given to the existential risk threat often depends on highly contestable ethical assumptions. To begin with, talk of existential risk is often elaborated within a utilitarian framework that reduces morality to the imperative to maximise overall welfare. But, as we have seen, this is a deeply problematic theory, not least due to its propensity to sacrifice the vital interests and rights of individuals on the altar of a supposed common good. Moreover, 'long-termist' proponents of existential risk make the further, deeply questionable assumption that in the process of utilitarian calculation, the lives of countless future people who might exist thousands of years hence are to be given equal weight to those of actually existing human beings in the here and now.[78]

[77] Rishi Sunak, "Rishi Sunak & Elon Musk: Talk AI, Tech & the Future", 2023, Available at: https://www.youtube.com/watch?v=R2meHtrO1n8&ab_channel=RishiSunak (Accessed 16 June 2024).
[78] For one proponent of this view, see William MacAskill, *What We Owe the Future* (Basic Books, 2022).

Another aspect of the discourse of existential risk is that it conceives of the emergence of AGI as the creation of a powerful set of tools in order to advance human purposes. But, especially if we correct for the unduly restrictive conception of 'intelligence' that is at play here, and consider what sort of beings really could replicate human cognitive functioning across its entire range, then it is highly likely that the beings in question would possess an intrinsic moral status, and perhaps be the bearers of rights. To this extent, to construe the problem raised by AGI as one of 'control', and the 'alignment' of 'instruments' or 'tools' with human purposes, seems deeply morally ill-judged, as if the issue were one of creating a new race of slaves.

The overall conclusion is that while it cannot be dismissed, the worry about existential risk is seriously over-hyped relative to the other, more concrete and imminent, challenges posed by AI, such as the way it perpetuates and deepens existing injustices, its potentially devastating impact on democracy, work opportunities, and the prospects for living in a world characterised by meaningful forms of human interaction.

In the event, the Bletchley declaration went far beyond 'safety' in its ordinary sense to enumerate an open-ended array of values, including human rights and the UN's Sustainable Development Goals.[79] So, no actual harm was done, you might think, by the potentially Procrustean safety framing. Perhaps 'safety', and the robot apocalypse scenario it signals, were effective rhetorical means of attracting the attention of the wider public and bringing to the summit table parties with disparate ideological perspectives. After all, who could oppose AI 'safety' with a straight face, whatever their political leanings? Certainly, one of the summit's biggest achievements in this regard was the inclusion of China, since there can be no effective global regulation of AI without China's involvement. And it may well be that the 'safety' framing helped facilitate China's participation, both from the perspective of the Chinese government itself and from that of Western states who might otherwise be hesitant about engaging China. But we should be careful not to confuse slogans that might have rhetorical or strategic value with the ultimate values our regulation of AI should serve.[80]

---

[79] John Tasioulas, Hélène Landemore, and Nigel Shadbolt, 'Bletchley Declaration: International Agreement on AI Safety is a Good Start, But Ordinary People Need a Say - Not Just Elites', *The Conversation* (Nov. 7, 2023) https://theconversation.com/bletchley-declaration-international-agreement-on-ai-safety-is-a-good-start-but-ordinary-people-need-a-say-not-just-elites-217042

[80] It is worth noting that for the next summit in the series - the AI Seoul Summit in May 20-21, 2024 - the word 'safety' was quietly dropped from the title. Meanwhile, the "International Scientific Report on the Safety of Advanced AI: Interim Report" published a few days before the Seoul gathering downplays the issue of existential risk and gives more emphasis to risks of bias, fake media, privacy violations, and economic dislocation. https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai

The Aristotelian Framework and Three 'Digital Empires'

We believe that the Aristotelian approach outlined in Parts I and II is superior to each of the rival ideological templates that the legal scholar Anu Bradford has identified as underlying the regulatory strategies of the world's three "digital empires" that are currently engaged in a struggle for global regulatory domination.[81] She calls these three templates "state-driven" (China), "market-driven" (United States), and "rights-driven" (European Union). Now, of course, this regulatory trichotomy is somewhat stylised and abstracts from important complexities, e.g. the presence of rights-based concerns in the US approach (e.g. the White House Office for Science and Technology Policy's 2022 Blueprint for an AI Bill of Rights[82] and the objective of fostering market-driven innovation in the EU's AI Act, etc.). But Bradford's tripartite division is nonetheless helpful in identifying the primary regulatory focus in each jurisdiction.

According to the Aristotelian approach developed here, each of these approaches is seriously flawed in its over-emphasis on one particular regulatory institution or consideration, but in each case this element needs to be integrated into a broader Aristotelian framework if it is to be properly understood and to function as it should. From an Aristotelian perspective, the first two approaches accord excessive significance to a particular kind of social institution, the state and the market respectively, while the third exaggerates the role of a particular kind of ethical consideration, that of individual rights.

Most obviously, the Chinese framework suffers from an authoritarianism that confers excessive power on the state over individuals, families, political associations, churches, business and other components of the political community. Politically, it does not embody anything approaching a defensible conception of free and equal citizens ruling and being ruled in turn which, as we have seen, necessitates a robustly participatory form of democracy. But the Aristotelian worry about authoritarianism extends far beyond the domain of political participation. It also applies to the state's obligation to respect the self-determination of individual citizens in their day-to-day lives by empowering them to make voluntary choices, often in the context of associations like families, businesses and political parties, in order to plot their course through life and, in the process, autonomously form and give expression to their individual characters. Such respect for individual self-determination is, for example, incompatible with the

[81] Anu Bradford, *Digital Empires: The Global Battle to Regulate Technology* (Oxford University Press, 2023).
[82] Office of Science and Technology Policy, "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People". Available at: https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Chinese government's 'social credit' policy for assigning entitlements to individuals and associations, a policy implemented in part by AI-driven surveillance tools.[83] Still, that the state has a vitally important role to play in advancing human flourishing is an important tenet of Aristotelian politics: It is the institution with the most comprehensive responsibility for furthering the common good of a political community, but one that needs to be exercised with due regard for the autonomy that should be accorded to individuals, families, corporations, and other associations.

The US market-driven framework, by contrast, is misguided to the extent that it accords excessive ethical significance to the promotion of wealth (measured by such indices as GDP) through the operation of the free market. Economic prosperity is a vital aspect of the common good on an Aristotelian view, but it is only one component of it. Wealth is itself not of intrinsic value, but valuable only as a means to securing things which themselves are of intrinsic value, such as health, understanding, justice, and so on. Moreover, focussing exclusively on the maximisation of wealth fails to attend to the significance of how wealth is distributed and, in particular, the risk that gross economic disparities will undermine the democratic ethos of freedom and equality. Alternatively, if the promotion of wealth is regarded as a mechanism for the utilitarian project of maximising the overall satisfaction of preferences, then the reply, as we have seen, is that there is no value in fulfilling preferences per se.

Of course, this still leaves an important role for economic prosperity, and for the free market (which must not be confused with actually existing corporate capitalism), in promoting the common good, just as the critique of authoritarianism leaves an important role for the state. The market is an invaluable mechanism for upholding the economic freedoms of individuals, creating incentives for investment and innovation, and channelling scarce resources to their most productive uses. But the market mechanism is not sufficient, not just because of 'market failures' in achieving economic efficiency, but because much more matters to us than efficiency. Moreover, it is arguable that the effective operation of the market, even in terms of economic efficiency, strongly presupposes not only other institutions (e.g. private property, a judicial system) but also other values, such as trust, honesty, and fairness.

Finally, the EU's rights-based approach is exemplified by its recent AI Act, which adopts a risk-based framework for the regulation of AI systems, with threats to human

[83] Zeyi Yang, "China just announced a new social credit law. Here's what it means." *MIT Technology Review* (2022). Available at: https://www.technologyreview.com/2022/11/22/1063605/china-announced-a-new-social-credit-law-what-does-it-mean/ (Accessed 16 June 2024).

rights being key determinants of risk.[84] Now, as we saw, there is controversy among interpreters of Aristotle regarding whether he, or ancient Greek philosophers generally, had the concept of a right, i.e. a moral entitlement owed to an individual as a matter of duty. But leaving this aside, we believe the notion of a human right, one possessed by all human beings simply in virtue of their humanity, can be developed within the Aristotelian framework, but that this has two highly consequential implications.[85]

The first is that rights are not ethically self-standing; although (contrary to the dominant European human rights doctrine of 'proportionality') they are not simply to be identified with any human interest, their justification as serious moral requirements depends on deeper ethical premises about the elements of human flourishing. In large part, human rights concern how the good of each and every human being imposes duties on others to respect and protect that good in various ways, e.g. by not torturing them or enslaving them, by providing them with educational and work opportunities. The second is that human rights do not exhaust the considerations that shape the regulation of AI, but have to be mobilised in tandem with other ethical concepts. These include virtues, such as civility, loyalty, toleration, and mercy, which typically are not exclusively concerned with respecting others' rights. We look for more, for example, from a carer of elderly people or a schoolteacher than that they will respect the rights of those in their charge. We also expect qualities such as friendliness, patience, kindness and so on. The relevant ethical concepts also include proper respect for non-human animals and the natural environment, which go beyond any anthropocentric concern with human interests and also beyond the individualistic focus of rights to encompass common goods (such as the common good of preserving a beautiful, pristine natural environment) to which no one may have a right.

In short, an Aristotelian approach can incorporate the elements of good sense embodied in statist, market-driven, and rights-driven approaches, while transcending their evident deficiencies when they are mistakenly treated as comprehensive frameworks for the regulation of AI and digital technology. To this extent, the Aristotelian framework provides a broad intellectual plateau that can facilitate genuine dialogue and mutual learning among the three major 'digital empires' as they experiment in AI regulation, enabling each of them to overcome the deficiencies

[84] For some concerns about gaps in the human rights protections afforded by the EU's AI Act, see J. Tasioulas and C. Green, 'The EU's AI Act at a Crossroads for Rights' *AI Ethics at Oxford Blog* Dec. 3, 2023 https://www.oxford-aiethics.ox.ac.uk/eus-ai-act-crossroads-rights

[85] The next paragraph summarises ideas more fully developed in John Tasioulas, "Saving Human Rights from Human Rights Law." *Vanderbilt Law Review* 52(5)(2019), pp. 1167-1207.

inherent in their one-sidedness. With its emphasis on a universal and pluralistic conception of human flourishing, the plateau could also accommodate other countries, especially in the Global South, that have been systematically excluded from a role in shaping the direction of AI's development. In the upshot, we are likely to discover that a great variety of regulatory approaches to AI are acceptable within the broad terms of the Aristotelian framework, and that distinct political communities should be given considerable leeway to shape their regulations in accordance with their own priorities, subject to the need for some overarching regulatory framework at the global level. It is to the difficult question of the global regulation of AI that we turn next.

Global Regulation

Does an Aristotelian approach to ethics and politics offer any resources for the kind of trans-national cooperation necessary for the global regulation of AI? The answer is yes, by extrapolation from Aristotle's understanding of interdependence and self-sufficiency. Aristotle's conception of the polis as a natural end was predicated on the fact of human interdependence, the need we humans have of one another for material existence (living), which is a precondition for moral flourishing (living well). As such, the natural impulse of the individual to pursue personal flourishing is necessarily linked to social existence. That linkage implies a concern for the flourishing of others; other-regardingness is an essential component of the rational pursuit of one's own well-being. The principle of interdependence drives the natural process of scaling up (*Politics* I.2): from the individual to the family, from families to villages/neighbourhoods, from villages/neighbourhoods to the polis. The Aristotelian individual was neither an atomised unit nor a metaphysically subordinate component of the state: The individual remains a natural "part" of the family and of the intermediate community of the village/neighbourhood. Each individual citizen owes a primary and rational allegiance (set of civic/social duties) to the state – such that the state can justly and reasonably ask the citizen to contribute part of his time and property to the collective welfare and risk his very existence in its defence.

For Aristotle, the polis, as a natural community with natural bounds (and therefore with clear physical borders, and a defined citizenship), concludes the process of natural scaling-up via interdependence. The polis is a natural end because it is necessary and (when rightly structured) sufficient as a social environment enabling the flourishing of each of its (citizen) members. It is the *smallest* possible self-sufficient (autarkic) community (*Politics* 1252b7-9, 27-30). Ideally, the polis would produce all it required

(*Politics* 1326b26-30), but that ideal was not attainable in practice. His prescription for "best-possible" autarky allowed for overseas trade (*Politics* 1327a25-27) while insisting on sovereign independence: the polis must not depend on external partners for sustaining the material or moral conditions of its own well-being. The polis was also the *largest* community whose members could know one another's virtues. The polis and its membership must be "readily taken in at a glance" (*eusunoptos*: *Politics* 1326b22-25). With the right knowledge of one another, citizens will be able to distribute goods justly and be self-governing according to the principle of "rule and be ruled in turn."

Aristotle's intuitive proof of the scale limit was Babylon, a city that cannot be a polis, due to its immense size, which precludes effectively communicating essential information among its residents (*Politics* 1276a27-30). But what if Aristotle were faced with a world in which *only* "Babylon-sized" states could be self-sufficient and in which a (potentially) flourishing state could be much larger than the "at a glance" limit because technological and institutional development allowed for effective communication and the relevant kinds of knowledge of its citizens of one another? Under such conditions, "rule and be ruled" could be scaled up to the size of the modern nation-state. This requires of course that institutional design and technology are up to the task. As noted above, the institutions of modern representative forms of democracy fall short of the Aristotelian demand for truly active participation in ruling, but technological advances in AI could at least potentially address that shortcoming.

What if, next, Aristotle were faced with a world in which self-sufficient and flourishing communities could no longer be sustained at the level of the bounded state, but required extensive cross-border cooperation, among a great many individuals, via cooperating states, at a global level? This requirement might operate in just the way that state-level self-sufficiency had required coordination among individuals via cooperating families and then villages/neighbourhoods. When self-sufficiency is not sustainable at the level of the bounded state, the fundamental interdependence principle drives the need for rational cooperation beyond the level of the state.

The questions then become: Could that need be met through inter-state cooperation? And, given our framework, would meeting that need break the bounds of Aristotelian forms of cooperation? We think that the answer to the latter question is 'no': Aristotle himself lived in an emerging Hellenistic world, one in which polis communities would exist under an 'umbrella' imperial regime. He understood the implications for the polis of the world being brought about by his sometime employers, Philip II and Alexander

III. Aristotle recognised that emergence as potentially valuable, insofar as it enabled the creation of new poleis – which might be structured as "best possible" states. He did not suppose that the emerging new order would end conflict between poleis (n.b. *Politics* book 7 on the necessity of strong city walls, his concerns about cross-border threats). But it did, arguably, open the way for thinking about some kind of supra-polis political order and what that order might undertake (other than extract taxes to sustain itself).

An cooperative international order aimed at global sustainability could (as the Hellenistic empires in fact did) leave intact the state as the locus of primary locus of civic duty, and leave states to be self-governing.[86] States would remain concerned with developing a "local" conception of the common good, a conception that need not be simply subsumed into the global good. So, we might imagine that a recognition of a global sustainability requirement is compatible with the possibility of continued competition and potentially even conflict between sovereign states. Cooperation at a global level on the conditions that enable living and living well, need not imply cooperation that eliminates the bounded community as the primary (although not unique) locus of duty.

The interdependence and self-sufficiency/sustainability principles entail a responsibility on the part of individuals and states to enact and enforce global rules that enable the material/environmental conditions that are the necessary conditions for human existence, and the moral conditions of flourishing. In the 21st century, this entails, at least, rules concerning the global environment (e.g. climate change) insofar as environmental degradation puts the conditions of mere life at risk. And it entails rules about AI, insofar as AI potentially puts the free exercise of human capacities of pro-social communication, the use of reason, and collective self-government by citizens at risk.

That gives us a possible Aristotelian justification for international law and institutions – i.e. rules that could legitimately bind individual states and justly trump their interests, in cases in which the behaviour of states (or individuals or organisations) threatened the sustainable material conditions of existence that make flourishing possible. And likewise, international rules could legitimately bind and trump the interests of states and corporations in developing and deploying AI if and when it threatens the legitimate

[86] John Ma, *Polis: A New History of the Ancient Greek City-State From the Early Iron Age to the End of Antiquity* (Princeton, 2024)

interests of other states in sovereign self-government or the moral flourishing of persons.

Given that AI seems certain to have a substantial impact on the lives of everyone, everywhere, AI is an area in which an Aristotelian approach suggests that international rules may come into play. Those who are not citizens of the rich and powerful countries in which AI is being developed, and who do not share in the decisions made by the rich and powerful corporations that are leading AI design, are at risk of losing (or further losing) the necessary level of local independence. They may fail to retain control of their own collective destinies and thereby lose the opportunity to live appropriately political lives. For example, the question of how people in the global South can be meaningful participants in how AI is deployed in their own communities, and how they will gain adequate access to the resources (hardware, software, training) – and the correlative benefits for flourishing – are therefore raised by and ought to be addressed with the Aristotelian framework.

An Aristotelian justification for international order would also set firm limits on the authority of international institutions. The authority of an international institution to bind states, organisations, or individuals would be restricted to cases in which the conditions of human material or moral well-being were being put at risk, and in which individual states were unable to act on their own initiative, with their own local resources, to rectify the situation. Those would include situations in which the relevant conditions could only be achieved and sustained by cooperation at a global level. Some level of environmental and AI regulation appears to us to fall within that ambit.

But the Aristotelian frame also highlights the importance of retaining state sovereignty, as well as local control of decision-making at sub-state levels. Per the principle of subsidiarity, salient decisions should be made as close as possible to the point at which they have their greatest effect. In sum, an Aristotelian approach to international regulation seeks a middle ground - a mean between highly intrusive inter-state institutions that are incompatible with robust state sovereignty on the one hand, and, on the other, conditions of international anarchy that threaten to deny many, perhaps even all people on the planet, the resources we need to live truly flourishing lives.

A Human Right to a Human Decision

There are many momentous issues that potentially fall within the scope of the global regulation of AI, which include but are not limited to: limiting or prohibiting the use of AI in military contexts, ensuring that the most powerful AI models are open source for the purposes of transparency and accessibility, mitigating the environmental impact of AI, and so on. Our white paper has sought not to make specific regulatory proposals, but rather to articulate a general ethical framework that can guide us in shaping and evaluating such proposals. Part of that framework is a key role for democratic publics whose deliberations and decision-making may be assisted but not preempted by the contributions of philosophers, technologists or experts of other kinds.

However, we now wish to conclude with one global regulatory proposal that resonates strongly both with the distinctive ethical challenges posed by AI and the distinctively humanistic preoccupations of the Aristotelian framework. This is the proposal that, in the age of AI, we need to recognise at least one novel human right: a human right to a human decision.[87] The general category of decisions we have in mind are those which bear significantly on the rights, duties, or basic interests of others. Illustrative cases include decisions to hire someone, to sentence a criminal to imprisonment, to determine eligibility to receive a loan or social welfare benefits, to admit to a university, to assign priority in the allocation of medical treatment, and so on.

Affirming a human right to a human decision involves the idea that with respect to certain decisions within this general class either (a) they should not be made by an AI system, or (b) if it is permissible for them to be made by an AI system, those potentially subject to its decisions should have the power either to (i) opt out of an AI-based decision-making process in favour of a human decision, or (ii) appeal from the AI-based decision to a human decision-maker. Moreover, the claim is not only that it is in some general sense wrong to deviate from these requirements, but rather that there is a *human right* to adherence to these requirements. This means that there is a moral right, on the part of each human being, that these requirements be complied with in their case.[88] In addition, this moral right furnishes a basis for incorporating such a

[87] The idea of such a right is gaining traction in various jurisdictions, including the EU and the US. Article 22 of the European Union's General Data Protection Regulation sets out a qualified 'right not to be subject to a decision based solely on automated processing'. https://gdpr-info.eu/ Meanwhile, the Blueprint for an AI Bill of Rights, published by the White House Office of Science and Technology Policy in 2022, includes the guideline that individuals should be able to 'opt out from automated systems in favour of a human alternative, where appropriate'. https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[88] John Tasioulas, 'A Human Right to a Human Decision' (forthcoming)

novel right into the corpus of international human rights law that can give effect to the background moral right.[89]

To assess the threat to human interests and dignity to which this proposed new right is a response, it is important to appreciate the arguments that might be made for delegating consequential decisions to AI systems. Here, the context of legal adjudication is instructive, so we focus on it for illustrative purposes.

Legal rights can mean very little if those upon whom they are conferred are unable to uphold them, including through access to the judicial system. Yet the worldwide situation regarding access to justice is bleak with the OECD estimating that only 46% of the world's population lives under the protection of the rule of law. Meanwhile the backlog of legal cases awaiting trial throughout the world includes 30 million cases in India and 100 million in Brazil.[90] Among the many possible ways that AI might help address this situation, some of the more enthusiastic proponents of AI-based justice have advocated the future deployment of AI systems to deliver binding legal decisions – in effect, AI systems as judges. The mooted benefits this would bring include (a) potentially huge efficiency gains, since these AI systems will be faster and cheaper than human judges, (b) potentially more accurate decisions, given that AI systems will not be vulnerable to human cognitive and other biases or vulnerabilities such as the need for sleep, as well as (c) more consistent decisions, ensuring that like cases are treated alike, in contrast to the 'noise' (unwanted variability) that notoriously afflicts legal decision-making by humans.[91]

Of course, advocates of this proposal readily concede that we are nowhere near the point at which any AI system can pass a 'legal Turing test', generating decisions that a panel of human legal experts would regard as indistinguishable in quality to those that a good human judge might deliver. But the claim is that this is a feasible goal and that there is no insuperable principled objection to deploying AI judges provided they have passed this test. Like many proposals to replace human endeavour with automated systems, this argument is heavily outcome-focussed. It promotes the use of AI on the basis of its potential to generate valuable outcomes (in this instance, correct legal

[89] Yuval Shany, 'A Case for a New Right to a Human Decision Under International Human Rights Law', https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4592244 (accessed June 14, 2024)
[90] Richard Susskind, *Online Courts and the Future of Justice* (Oxford University Press, 2019), ch. 28.
[91] See, for example, Eugene Volokh, 'Chief Justice Robots', *Duke Law Journal* 68 (2019) 1,135 and Richard Susskind, *Online Courts and the Future of Justice* (Oxford University Press, 2019), ch. 28. For a somewhat more moderate take, largely conditioned by considerations of feasibility in light of human resistance to algorithmic decision-making, see Daniel Kahneman, Olivier Sibony, and Cass Sunstein, *Noise: A Flaw in Human* Judgment (London, William Collins, 2021).

decisions). Insofar as it addresses the *process* through which these outcomes are generated, it focuses overwhelmingly on its efficiency as compared with human decision-making.

Now, to a large extent, this case can be questioned on its own terrain of outcomes. Will an AI system deliver correct decisions, or will it be subverted by biases in the data on which its algorithm was trained or biases inadvertently instilled into it by its designers? Again, the emphasis on 'noise' reduction may be insensitive to the fact that a plurality of conflicting decisions are eligible options in a given case. After all, if plural and conflicting values are implicated in a judicial decision-making, such as the need to balance justice and mercy, there is no reason to suppose there is one optimal way of striking this balance, hence one optimally correct sentence.[92] Here, there is a looming risk that the tools at our disposal come to distort the nature of the problems that confront us, falsely converting them into readily quantifiable exercises in optimization.

Moreover, looking to outcomes beyond the quality of decision, the widespread deployment of AI adjudicative tools risks having unwelcome side-effects. Their use may lead to the atrophying of judicial virtues among humans who will have been deprived of the opportunity to develop their capacity for legal judgement in the context of real-life decision-making rather than the law school classroom. This in turn might erode the level of legal competency we can deploy to subject AI adjudicative systems, and the companies that produce them, to effective democratic scrutiny. More generally, the increased automation of the judicial system might foster an even deeper sense of disillusionment and alienation on the part of most people towards the legal system, casting it as a domain in which humans mostly figure as mere passive consumers.

All these considerations may play a part in defending a right to a human decision, but they are largely hostage to the quality of the decisions produced by AI adjudicative tools, which may in fact improve markedly. The core case for a right to a human decision, however, will also invoke values related to the *process* of decision-making, not just its outcome. We have already encountered such process-based considerations

[92] For the implications of value pluralism and incommensurability for automated decision-making, see John Tasioulas 'Artificial Intelligence, Humanistic Ethics', *Daedalus* (2022) 151(2): 232-243 https://www.amacad.org/sites/default/files/daedalus/downloads/Daedalus_Sp22_AI%20%26%20Society_1.pdf; and Ruth Chang, 'Human in the Loop!', in David Edmonds (ed.), *AI Morality* (OUP, forthcoming). For an attempt to develop an AI system, Kaleido, that is responsive to the plurality of values in engaging with moral questions, see T. Sorensen, L. Jiang, J. Hwang, S. Levine, V. Pyatkin, P. West, N. Dziri, X. Lu, K. Rao, C. Bhagavatula, M. Sap, J. Tasioulas, Y. Choi, 'Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties', arXiv preprint arXiv:2309.00779 (2023/9/2).

in our discussions of work and democracy in Part II. Those activities are valued not just because of the valuable outcomes they produce e.g. income or good law, but also because of the intrinsic value of the processes through which these outcomes are achieved. These are processes exemplifying achievement, for example, or respectful collective deliberation and decision-making about the common good among free and equal citizens. Similarly, in the case of legal adjudication, we value not only legally sound decisions and efficient processes, but processes that exhibit certain other intrinsic values that AI systems are not well-placed to exemplify. Here we mention just three: explainability, accountability, and solidarity.[93]

To begin with, we seek not just a correct legal decision, but also an *explanation* for it, one that provides a causally effective justification for it. By contrast, the workings of machine learning algorithms are often opaque, even to their designers, due to the potentially astronomical number of statistical patterns among vast amounts of data that may be involved. And even when they reach the correct result, they may do so by means of a route that does not connect the decision with the good reasons for it. Consider the AI system Lex Machina that can successfully predict the outcome of patent litigation at a level comparable to top patent lawyers, but uses justificatorily 'irrelevant' factors such as the monetary value of the claims and the names of the judge and the litigants as bases for prediction.[94] Litigants rightly desire not just a correct decision but a justification for it that is not just an *ex post* rationalisation but the operative cause of the decision rendered by the judge. As we saw in Part I, it is of the essence of Aristotelian virtue that the right thing is done for the sake of the reasons that make it right, and out of a settled disposition to act for those reasons.

Even if an AI system were developed that could provide a causally efficacious justification for its decisions, we are still far away from any such system being *accountable* for its decisions. All such systems currently known to us, and likely to emerge in the foreseeable future, lack the capacity to exercise rational autonomy in choosing to apply the law and to affirm a specific decision in a particular case. Even the 'super-intelligent' AI systems foreshadowed by thinkers like Stuart Russell lack this capacity. Yet, in legal adjudication, we generally rightly desire not only a legally sound resolution of our case but also, and independently of that, a decision-maker who can

[93] The ensuing paragraphs develop ideas more fully explored in J Tasioulas, 'The Rule of Algorithm and the Rule of Law', in C. Bezemek, M. Potacs, A. Somek (eds), *Vienna Lectures on Legal Philosophy vol 3: Legal Reasoning* (Hart Publications, 2023), pp.17-39 and John Tasioulas, 'A Human Right to a Human Decision' (forthcoming). The former article employs Plato's comparison of free and slave doctors in *The Laws* 720b-e as a means of exploring these ideas.
[94] Richard Susskind, *Online Courts and the Future of Justice* (Oxford University Press, 2019), p.282.

be properly held accountable for that resolution. Of course, the humans who decided to deploy the AI adjudicative system will still be accountable for doing so, but this diffuse and distal accountability is not the same as the more immediate form of accountability that is in play when a human judge reaches a decision after consideration of one's own specific case.

It might be objected that the argument so far overlooks the possibility of real AGI, one that replicates human capacities across the board, that possesses consciousness and rational autonomy. But even if we contemplate this far-fetched possibility, there would remain a loss in *solidarity* in deploying AI judges. As we emphasised in Part I, humans are not simply autonomous rational agents but also creatures with the capacities, limitations, and vulnerabilities of a certain kind of biological and psychological nature. That nature is integral to the content of the reasons that apply to us. Not sharing that condition, AI systems would be outsiders to our outlook, even if they might sympathetically seek to engage with it, as we do in the case of non-human animals.

In the specific case of legal adjudication, the litigant would not face their judge on the plateau of a common humanity, with the sort of mutual understanding and give-and-take of reasons in dialogue that it makes possible. In showing mercy to another, for example, on the grounds that they have been the victim of an abusive upbringing or grossly unjust economic deprivation, there is a charitable response to a fellow human being, one grounded in an empathetic sentiment of 'there but for the grace of God go I'.[95] An AI judge, even one programmed to generate suitably merciful decisions in such cases, would not be fully able to participate in that empathetic sentiment, since they do not inhabit the shared human condition that it presupposes. Exactly the same merciful sentence will have a different significance depending on whether it is passed by a human or an AI judge. In the case of the latter, it cannot convey an empathetic response to the challenges that afflict those with a common human nature.

None of the foregoing considerations amount to a conclusive defence of a novel right to a human decision or lead to a tolerably precise account of the decisions to which it applies. A fuller specification of the content of the right will require the participation of democratic publics assisted by experts in law, medicine, education, and other decision-making domains. Moreover, there is no good reason to suppose that there is a unique class of decisions that falls under this right; as with the specification of other

[95] John Tasioulas, 'Mercy', *Proceedings of the Aristotelian Society* 103 (2003): 101-132.

human rights, there may be considerable room for legitimate cultural variation as to the content of the right to a human decision. Finally, nothing we have said here amounts to a case for the blanket exclusion of AI judges in all cases, although in general we will believe that AI adjudicative systems should primarily be used as tools that assist, rather than replace, human judges.

As always, the overarching objective remains that of integrating AI systems into human life in ways that enhance our capacities for rational self-direction as individuals and communities, thereby promoting both individual flourishing and the common good. In this endeavour, a novel and suitably qualified human right to a human decision can be a bulwark - one among many that we need - against all the forces that seek to use AI technologies in ways that disempower and diminish humanity.

Conclusion

The AI revolution through which we are living holds great promise to enhance individual and collective flourishing: delivering us from the drudgery of dangerous and unrewarding work, accelerating the process of scientific discovery, making access to education, health, and justice more efficient and widespread, and democratic citizenship more meaningful. But for this promise to be realised, and to ward off the dangers that are attendant upon this transformative technology, we need a compelling ethical framework to guide our choices about the development and deployment of AI systems. This framework, we have argued, does not need to be built from scratch. Rather, its key elements already exist in the ethical thought of Aristotle, a thinker who stood at the pinnacle of both scientific and philosophical achievement in his time.

The Aristotelian framework offers a truly humanistic or human-centred approach to the ethics of AI, one that makes the cultivation and exercise of our distinctive human capacities for reason, communication, and social engagement central to individual well-being and the common good. Moreover, it places great emphasis on a radically participatory conception of democracy in translating our ethical deliberations into political decisions. The Aristotelian framework is richer and more compelling than the two ethical approaches that have dominated the ethical discourse on AI in the West: variants of preference-maximising utilitarianism (or their economistic wealth-maximising cousins), on the one hand, and the legalistic discourse of human rights, on the other. These dominant alternatives operate with an impoverished scheme of values and tend to reduce AI ethics to a technocratic exercise, rather than an inclusive process of deliberation and decision-making oriented towards the common good.

Our core Aristotelian thesis is that AI systems should be conceived primarily as ‘intelligent tools’ that we can deploy to enhance the prospects of human flourishing, both individual and communal. It is this goal, not Artificial General Intelligence that replicates human cognitive capabilities across their entire spectrum, that should guide us in developing this powerful new technology. We need AI tools to enhance our ability to engage in meaningful and productive work, not to replace us in familiar human endeavours that are vital sources of achievement, friendship, and self-esteem. And alongside legitimate worries about the misuse of AI technology to subvert the informed, engaged, and civil public discourse that democracy requires, we should use it to build

participatory democracy in modern states. AI should enable us to realise the Aristotelian ideal of free and equal citizens, ruling and being ruled in turn.

No ethical framework can provide ready answers to the many highly specific and context-dependent regulatory challenges posed by AI. These must be the province of democratic publics informed (but not dominated) by technical and other experts. Yet the Aristotelian framework offers valuable markers for regulation: the need to articulate a convincing account of the positive good that AI can help us achieve under a humanistic and pluralistic account of our values; the limits of 'safety' as an overarching rubric for regulating AI; the need for global regulation of AI that remains appropriately targeted at those challenges that genuinely require universal standards and that does not usurp powers of decision that properly belong to states or local associations; and the importance of recognising a human right to a human decision to ensure that the most consequential decisions bearing on human interests, rights, and duties are not offloaded to systems that – for the foreseeable future, anyway – exhibit severe limitations in explainability, accountability, and human solidarity.

The Aristotelian framework is not a panacea. No philosophical framework could reasonably claim to meet all the profound ethical challenges raised by AI. But it does offer rational grounds for hope in that endeavour. The essential starting-point is the recognition that the evolution of AI technology is not a fate that we are helpless to affect. We have choices, and the real question is who gets to make them, how, and on what basis. What is needed, above all, is informed and engaged democratic citizens, motivated to deliberate as free and equal persons about the place of AI technology in achieving common goods. Our hope is that this paper has shown why and how each citizen can play their part in choosing our shared future.[96]

[96] The authors wish to acknowledge the helpful feedback on earlier versions of this paper from Rahul Santhanam and the members of their graduate class on Ethics, Democracy, and Technology held at the University of Oxford in Trinity Term, 2024. They are especially grateful to Kyle van Oosterum for invaluable and timely editorial assistance..

About the Authors

Josiah Ober is the Constantine Mitsotakis Chair in the School of Humanities and Sciences at Stanford University, and currently Eastman Visiting Professor at Balliol College, Oxford. He specialises in the areas of ancient and modern political theory and historical institutionalism. His primary appointment is in Political Science; he holds a secondary appointment in the Classics and courtesy appointments in Philosophy and the Hoover Institution. His most recent books are *The Greeks and the Rational: The Discovery of Practical Reason* (University of California Press, 2022), *Demopolis: Democracy Before Liberalism in Theory and Practice* (Cambridge University Press, 2017) and (with Brook Manville) *The Civic Bargain: How Democracy Survives* (Princeton University Press 2023). His ongoing work focuses on rationality (ancient and modern), the theory and practice of democracy, and the politics of knowledge and innovation. Recent articles and working papers address AI ethics, socio-political systems, economic growth and inequality, the relationship between democracy and dignity, and the aggregation of expertise. He is author or co-author of about 100 articles and chapters and several other books, including *Fortress Attica* (1985), *Mass and Elite in Democratic Athens* (1989), *The Athenian Revolution* (1996), *Political Dissent in Democratic Athens* (1998), *Athenian Legacies* (2005), *Democracy and Knowledge* (2008), and *The Rise and Fall of Classical Greece* (2015).

John Tasioulas is the inaugural Director of the Institute for Ethics and AI and Professor of Ethics and Legal Philosophy, University of Oxford. He was previously Yeoh Professor of Politics, Philosophy & Law at King's College London and Quain Chair of Jurisprudence at University College London. Professor Tasioulas has held visiting positions at the Australian National University, the University of Chicago, Harvard University, the University of Melbourne, and the University of Notre Dame. He is currently a Senior Fellow in Schmidt Sciences' AI2050 programme. He has also acted as a consultant on human rights for the World Bank and served as a member of the International Advisory Board of the European Parliament's Panel for the Future of Science and Technology (STOA) and the Greek Prime Minister's High-Level Advisory Committee on AI. He has published widely in moral, legal, and political philosophy. His recent work addresses topics such as the rule of law, human rights, and a humanistic ethics of AI. He is the author of *On Justice and Mercy: Essays in Moral and Legal Philosophy* (Oxford University Press, forthcoming) and the co-editor of *The Philosophy of International Law* (Oxford University Press, 2010) and the editor of *The Cambridge Companion to the Philosophy of Law* (Cambridge University Press, 2020).