AI ethics with Aristotle

By Josiah Ober and John Tasioulas

"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, Hayayo Miyazaki, in a video clip that was posted recently on the social media platform X. Earlier in the clip Miyazaki was shown sitting across a conference table with a group of chastened-looking technologists. 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".

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 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 placed into dialogue with the dominant ideological frameworks that are currently steering the development of AI technologies - ideologies centred on the promotion of economic growth, maximizing 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.

At the deepest level, the Aristotelian framework for thinking about AI consists of three core interlocking ideas:

(1) that human beings possess a distinctive nature as rational, communicative, and social animals, a nature that is not shared by existing AI systems or any such systems that might 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 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, 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.

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 instruments" 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), or of systematically replacing 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).

The Aristotelian framework is, we believe, superior to the world's three dominant regulatory approaches to digital regulation described by Anu Bradford in *Digital Empires:The Global Battle to Regulate Technology* (OUP, 2023) - 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 properly integrated. Moreover, Aristotle's concern with human interdependence and the conditions necessary for self-sufficiency provide the basis for an argument for international cooperation in the global regulation of AI. And it does so while preserving significant autonomy for distinct political communities within the architecture of global AI regulation, thereby guarding against global governance overreach.

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. And no general framework, however sound, can immunize 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.