Artificial Intelligence, Humanistic Ethics

John Tasioulas

Ethics is concerned with what it is to live a flourishing life and what it is we morally owe to others. The optimizing mindset prevalent among computer scientists and economists, among other powerful actors, has led to an approach focused on maximizing the fulfilment of human preferences, an approach that has acquired considerable influence in the ethics of AI. But this preference-based utilitarianism is open to serious objections. This essay sketches an alternative, "humanistic" ethics for AI that is sensitive to aspects of human engagement with the ethical often missed by the dominant approach. Three elements of this humanistic approach are outlined: its commitment to a plurality of values, its stress on the importance of the procedures we adopt, not just the outcomes they yield, and the centrality it accords to individual and collective participation in our understanding of human well-being and morality. The essay concludes with thoughts on how the prospect of artificial general intelligence bears on this humanistic outlook.

Ethics is, first and foremost, a domain of ordinary human thought, not a specialist academic discipline. It presupposes the existence of human choices that can be appraised by reference to a distinctive range of values. The delimitation of this range, among other values such as aesthetic or religious values, is philosophically controversial. But on a fairly standard reading, two very general, interlocking questions lie at the heart of ethics: What is it to live a good or flourishing life? And what is it that we owe to others, notably fellow human beings, but also nonhuman animals or even inanimate nature? The first question brings us into the territory of individual well-being; the second into that of morality, especially the obligations we owe to others and the rights they hold against us. Philosophers expound theories of well-being and morality and their interrelations, but all of us, in living our lives, constantly make choices that reflect answers to these questions, however inchoate or unconscious they may be.

Engagement with ethics is inescapable in decision-making about artificial intelligence.[1] The choices we make regarding the development and deployment of AI-based technologies are ultimately intelligible only in terms of the fallible pursuit of ethical values such as the acquisition of knowledge and control or the promotion of health, justice, and security. Moreover, all forms of "regulation" that might be proposed for AI, whether voluntary self-regulation in deciding whether

https://doi.org/10.1162/DAED_a_01912

to use a social robot as a caregiver, or the social and legal norms that should govern the manufacturing and use of such robots, ultimately implicate choices that reflect judgments about ethical values and their prioritization.

A clear-eyed appreciation of the pervasive significance of ethics for AI is sometimes obscured by an odd contraction that the idea of ethics is liable to undergo in this domain. So, for example, Kate Crawford, author and founder of the AI Now Institute, urges us to "focus less on ethics and more on power" because "AI is invariably designed to amplify and reproduce the forms of power it has been deployed to optimize."[2] But what would the recommended focus on power entail? For Crawford, it means interrogating the institutional power structures in which AI is embedded by reference to ideas of equality, justice, and democracy. But the irony is that these three ideas are either themselves core ethical values or, in the case of democracy, need to be explicated and defended in terms of such values.

Nonetheless, Crawford's injunction usefully prompts reflection on the various ways the idea of ethics has been unduly diminished in recent discussions about AI, no doubt partly a result of the prominent role of big tech players in shaping the field of "AI ethics" to limit the threat it poses to their commercial ambitions. Consider three ways the diminishment of ethics is typically effected.

Content. The content of ethical standards is often interpreted as exclusively a matter of fairness, which is primarily taken to be a relational concern with how some people are treated compared with others. Illustrations of AI-based technology that raise fairness concerns include facial recognition technology that systematically disadvantages darker-skinned people or automated resume screening tools that are biased against women because the respective algorithms were trained on data sets that are demographically unrepresentative or that reflect historically sexist hiring practices. "Algorithmic unfairness" is a vitally important matter, especially when it exacerbates the condition of members of already unjustly disadvantaged groups. But this should not obscure the fact that ethics also encompasses nonrelational concerns such as whether, for example, facial recognition technology should be deployed at all in light of privacy rights or whether it is disrespectful to job applicants in general to rank their resumes by means of an automated process.[3]

Scope of application. Ethics is sometimes construed as narrowly individualistic in focus: that is, as being concerned with guiding individuals' personal conduct, rather than also bearing on the larger institutional and social settings in which their decisions are made and enacted.[4] In reality, however, almost all key ethical values, such as justice and charity, have profound implications for institutions and patterns of social organization. Plato's *Republic*, after all, sought to understand justice in the individual soul by considering it "writ large" in the polity. Admittedly, some philosophers treat political justice as radically discontinuous from justice in the soul. The most influential proponent of the discontinuity thesis in

recent decades is John Rawls, who contends that pervasive reasonable disagreement on ethical truth disqualifies beliefs about such truths from figuring as premises in political justification.[5] This is a sophisticated controversy, which cannot be addressed here, save to note that this kind of move will always face the response that the phenomenon of reasonable disagreement, and the need for respect that it highlights, is itself yet a further topic for ethical appraisal, and hence cannot displace the need to take a stand on ethical truth.[6]

Means of enforcement. There is a widespread assumption that ethics relates to norms that are not properly enforceable – for example, through legal mechanisms – but instead are backed up primarily by the sanction of individual conscience and informal public opinion. But the general restriction of ethics to "soft" forms of regulation in this way is arbitrary. The very question whether to enact a law or other regulatory norm and, if so, how best to implement and enforce it, is one on which ethical values such as justice and personal autonomy have a significant bearing. Indeed, there is a long-standing tradition, cutting across ideological boundaries, that identifies justice precisely with those moral rights that should in principle receive social and legal enforcement.

In short, we should reclaim a broad and foundational understanding of ethics in the AI domain, one that potentially encompasses deliberation about any form of regulation, from personal self-regulation to legal regulation, and which potentially has radical implications for the reordering of social power.

Given its inescapability, ethical thought is hardly absent from current discussions around AI. However, these discussions often suffer from a tendency either to leave inexplicit their operative ethical assumptions or else to rely upon them uncritically even when they are made explicit. We can go even further and identify a dominant, or at least a prominent, approach to ethics that is widely congenial to powerful scientific, economic, and governmental actors in the AI field.

Like anyone else, AI scientists are prone to the illusion that the intellectual methods at their disposal have a far greater problem-solving purchase than is warranted. This is a phenomenon that Plato diagnosed in relation to the technical experts of his day, artisans such as cobblers and shipbuilders. The mindset of scientists working in AI tends to be data-driven, it places great emphasis on optimization as the core operation of rationality, and it prioritizes formal and quantitative techniques. Given this intellectual orientation, it is little wonder that an eminent AI scientist, like Stuart Russell, in his recent book *Human Compatible: AI and the Problem of Control,* is drawn to preference-based utilitarianism as his overarching ethical standpoint.[7]

Russell's book takes the familiar worry that AI – in the form of an artificial general intelligence (AGI) that surpasses human intellectual capabilities – will even-

tually spiral out of control, unconstrained by human morality, with disastrous consequences. But what is human morality? Russell appears to take it as axiomatic that the morally right thing to do is whatever will maximize the fulfilment of human preferences.[8] In terms of our two core concerns of ethics, the fulfilment of human preferences is taken to encompass well-being, and the fundamental moral injunction is to maximize overall well-being thus conceived. So ethics is reduced to an exercise in prediction and optimization: which act or policy is likely to lead to the optimal fulfilment of human preferences?

But this view of ethics is notoriously open to multiple serious – I believe, fatal – objections. Its concern with aggregating preferences threatens to override important rights that erect strong barriers to what can be done to individuals. Why not feed a few Christians to the lions if their preferences to stay alive are outweighed by the preferences of a sufficiently large number of blood-thirsty Roman spectators? And that is even before we observe that many preferences are infected with racism, sexism, or other prejudices; that they may reflect false or incomplete information; or that they may be psychological adaptations to oppressive circumstances. Ethics operates in the crucial space of reflection on what our preferences should be, a vital consideration that makes a belated appearance in the last few pages of Russell's book.[9] It cannot take those preferences as ultimate determinants of value.

There are moral philosophers who defend versions of preference utilitarianism that are patched-up to address these difficulties. But the idea that preference utilitarianism is a highly contestable moral theory does not really register in Russell's book, which conforms with my suspicion that it approximates to a default position among leading actors in the AI field.

The same broad approach is heavily influential among leading economic and governmental actors. This is perhaps less obvious, since the doctrine is standardly modified by positing wealth-maximization as the more readily measurable proxy for preference satisfaction. Hence the tendency of GDP to hijack governmental decision-making around economically consequential technologies, with the resultant sidelining of values that are not readily catered to by the market, such as public goods like access to justice and health care or the preservation of a sustainable environment. Hence, also, the legitimation of profit maximization by corporations as the most effective institutional means to societal wealth maximization. Of course, many who adopt such an approach have never heard of utilitarianism or, if they have, may explicitly reject it. But one revealing indication of the dominance of an ideology is the way that people who disavow it can nonetheless remain in its intellectual grip.

A key priority for those working in the field of AI ethics is to elaborate an ethical approach that transcends the limitations and distortions of this dominant ethical paradigm. In my view, such a humanistic ethics – one

that encompasses aspects of human engagement with the ethical that are not adequately captured by the methods of natural science and mainstream economics, but that are the traditional concern of the arts and humanities – would possess at least the following three, interrelated features (the three Ps).

Pluralism. The approach would emphasize the plurality of values, both in terms of the elements of human well-being (such as achievement, understanding, friendship, and play) and the core components of morality (such as justice, fairness, charity, and the common good). This pluralism of values abandons the comforting notion that the key to the ethics of AI will be found in a single master concept, such as trustworthiness or human rights. How could human rights be the comprehensive ethical framework for AI when, for example, AI has a serious environmental impact beyond its bearing on anthropocentric concerns? And what of those important values to which we do not have a right, such as mercy or solidarity? Nor can trustworthiness be the master value. Being parasitic on compliance with more basic values, trustworthiness cannot itself displace those values.

Beyond the pluralism of values is their incommensurability. We are often confronted with practical problems that implicate an array of values that pull in different directions. In such cases, although some decisions will be superior to others, there may be no single decision that is optimal: in choosing an occupation, teaching may be a better field for me than surgery, but we cannot assume there is a single profession that is, all things considered, best, rather than a limited array of eligible alternatives that are no worse than the others. This incommensurability calls into question the availability of some optimizing function that determines the single option that is, all things considered, most beneficial or morally right, the quest for which has animated a lot of utilitarian thinking in ethics.

It is worth observing that confidence about the deployment of AI to minimize "noise" in human judgment – the unwanted variability, for example, in hiring decisions by employers or sentencing by judges – displayed in the important new work of Daniel Kahneman, Olivier Sibony, and Cass Sunstein, sometimes involves an implicit reductionism about the values at stake that downplays the scope for incommensurability.[10] For example, the authors treat bail decisions fundamentally as predictions of the likelihood that the accused will abscond or reoffend, sidelining considerations such as the gravity of the offense with which they have been charged or the impact of detention on the accused's dependents.[11] But such decisions typically address multivalue problems, and there is no guarantee that there is a single best way of reconciling the competing values in each case. This means not only that algorithms will need to be more sophisticated to balance multiple salient values in reaching a correct decision, but that much of what looks like noise may be acceptable variability of judgments *within* the range of rationally eligible alternatives.

Procedures, not only outcomes. Of course, we want AI to achieve valuable social goals, such as improving access to education, justice, and health care in an effective and efficient way. The COVID-19 pandemic has cast into sharp relief the question of what outcomes AI is being used to pursue: for example, is it enabling physicians to diagnose and triage patients faster and more effectively, or is it primarily engaged in profit-making activities, like vacuuming up people's attention online, that have little or no redeeming social value?[12] The second feature of a humanistic approach to ethics emphasizes that what we rightly care about is not just the value of the outcomes that AI applications can be used to deliver, but the procedures through which it does so.

If, for example, important practical decisions exhibit the phenomenon of incommensurability, then we may have good reason to ensure that they are assigned to humans, rather than to automated processes, to preserve a valuable form of autonomy for humans as they express and develop their tastes and characters in choosing from divergent, but rationally eligible, pathways in life. Of course, there is the further question of how to balance such autonomy against demands for consistency (or "noiselessness"), especially in public decision-making. Should we tolerate significant divergence in sentencing across judges, or should the demands for "horizontal equity" prevail, ensuring that like cases are treated alike? Proponents of the latter view often recommend the use of algorithms to guide or replace human decision-making. This itself is a difficult question of striking a balance between competing considerations in our legal culture, with no *ex ante* guarantee that one solution will emerge as superior overall.

But the case for according ultimate decision-making authority to humans can also be made even if we suppose that a single correct answer is always available. Take, for example, the use of AI in cancer diagnosis and its use in the sentencing of criminals. Intuitively, the two cases seem to exhibit a difference in the comparative valuing of the soundness of the eventual decision or diagnosis and the process through which it is reached. When it comes to cancer, generating the most accurate diagnosis may be all-important, it being largely a matter of indifference whether this is generated by an AI diagnostic tool or the exercise of human judgment. In criminal sentencing, however, being sentenced by a robot judge – even if the sentence is likely to be less biased or less "noisy" than one rendered by a human counterpart – appears to sacrifice important values, such as the ideal of reciprocity among fellow citizens that is central to the rule of law.[13]

This last point is familiar, of course, in relation to such process values as transparency, procedural fairness, and explainability. Even if the procedure followed by the judicial algorithm can be made transparent, there is a serious question – given, for example, the vast discrepancy between machine learning and ordinary human reasoning processes – whether it affords an explanation of the right kind, an explanation that a criminal defendant can grasp as offering intelligible reasons for

the decision to imprison him. But the point goes beyond the important issue of explainability. How does it feel to contemplate the prospect of a world in which judgments that bear on our deepest interests and moral standing have, at least as their proximate decision-makers, autonomous machines that do not have a share in human solidarity and cannot be held accountable for their decisions in the way that a human judge can?

Participation. The third feature relates to the importance of participation in the process of decision-making with respect to AI, whether as an individual or as part of a group of self-governing democratic citizens. At the level of individual well-being, this takes the focus away from theories that equate human well-being with an end state such as pleasure or preference-satisfaction. These end states could in principle be brought about through a process in which the person who enjoys them is passive: for example, by the government putting a happiness drug into the water supply. Contrary to this passive view, it would stress that successful engagement with valuable pursuits is at the core of human well-being.[14]

If the conception of human well-being that emerges is deeply participatory, then this bears heavily on the delegation of decision-making power to AI applications. One of the most important sites of participation in constructing a good life, in modern societies, is the workplace.[15] According to a McKinsey study, around 30 percent of all work activities in 60 percent of occupations could one day be automated.[16] Can we accept the idea that the large-scale elimination of job opportunities can be compensated for by the benefits that automation makes available? The answer partly depends on whether the participatory self-fulfilment of work can, any time soon and for the vast majority of those rendered jobless, be feasibly replaced by other activities, such as art, friendship, play, or religion. If it cannot, addressing the problem with a mechanism like a universal basic income, which involves the passive receipt of a benefit, will hardly suffice. Instead, much greater attention will need to be paid to how AI can be integrated into productive practices in ways that do not so much replace human work as enhance its quality, making it more productive, fulfilling, and challenging, while also less dangerous, repetitive, and lacking in meaning.[17]

Similarly, we value citizen participation as part of collective democratic self-government. And we do so not just because of the instrumental benefits of democratic decision-making in generating superior decisions by harnessing cognitive diversity, but also because of the way in which participatory decision-making processes affirm the status of citizens as free and equal members of the community.[18] This is an essential plank in the defense against the tendency of AI technology to be co-opted by technocratic modes of decision-making that erode democratic values by seeking to convert matters of political judgment into questions of technical expertise.[19]

At present, much of the culture in which AI is embedded is distinctly technocratic, with decisions about the "values" encoded in AI applications being taken by corporate, bureaucratic, or political elites, often largely insulated from meaningful democratic control. Indeed, a small group of tech giants accounts for the lion's share of investment in AI research, dictating its overall direction and setting the prevalent moral tone. Meanwhile, AI-enabled social media risks eroding the quality of public deliberation that a genuine democracy needs, such as by promoting the spread of disinformation, aggravating political polarization, or using bots in astroturfing campaigns. Similarly, the use of AI as part of corporate and governmental efforts to monitor and manipulate individuals undermines privacy and threatens the exercise of basic liberties, effectively discouraging citizen participation in democratic politics.[20]

As with workplace participation, we need to reflect seriously on how AI and digital technology more generally can enable, rather than hinder and distort, democratic participation.[21] This is especially urgent given the declining faith in democracy across the globe in recent years, including in long-established democracies such as the United Kingdom and the United States. Indeed, the disillusionment is such that, in a recent poll, 51 percent of Europeans favored replacing at least some of their parliamentarians with AI.[22] There is still time to salvage the democratic ideal that an essential part of civic dignity is participation in self-government.

An additional complexity here concerns how these two modes of participation – in the workplace and in politics – are connected. It is obvious that active participation in the two domains is mutually reinforcing in important ways. Thus, powers of reason and sociability that are developed in a participatory workplace, and that foster a sense of equal civic dignity, can be brought to bear in democratic deliberation about political questions, just as democratic control over the impact of new technologies on the workplace can help preserve and enhance its vital role as a site of genuine human fulfilment.[23]

I have mainly focused on narrow AI, conceived as AI-powered technology that can perform limited tasks (such as facial recognition or medical diagnosis) that typically require intelligence when performed by humans. This is partly because serious doubt surrounds the likelihood of artificial general intelligence emerging within any realistically foreseeable time frame, partly because the operative notion of "intelligence" in discussions of AGI is problematic,[24] and partly because a focus on AGI often distracts us from the more immediate questions of narrow AI.[25]

With these caveats in place, however, one can admit that thought experiments about AGI can help bring into focus two questions fundamental to any humanistic ethic: What is the ultimate source of human dignity, understood as the inherent value attaching to each and every human being? And how can we relate hu-

man dignity to the value inhering in nonhuman beings? Toward the end of Kazuo Ishiguro's novel *Klara and the Sun*, the eponymous narrator, an "Artificial Friend," speculates that human dignity – the "human heart" that "makes each of us special and individual" – has its source not in something within us, but in the love of others for us.[26] But a threat of circularity looms for this boot-strapping humanism, for how can the love of others endow us with value unless those others already have value? Moreover, if the source of human dignity is contingent on the varying attitudes of others, how can it apply equally to every human being? Are the unloved bereft of the "human heart"?

Questions like these explain the tendency among some to interpret the inherent value of each individual human being as arising from the special love that a supremely good transcendent being – God, represented by the sun, in Ishiguro's novel, which the solar-powered Klara treats as a kind of life-sustaining divinity – has for each human being in equal measure.[27] But invoking a divine being to underwrite human dignity leads us into obvious metaphysical and ethical quagmires, which in turn raise the difficult question of whether the inherent worth of human beings can be explicated within a broadly naturalistic framework.[28] Supposing that it can be, this is compatible with a distinct kind of dignity also inhering in other beings, such as nonhuman animals.

We are still struggling to integrate the value of nonhuman animals within our ethical thought. Doing so requires overcoming the baleful influence of longstanding practices in which animals are treated either as possessing merely instrumental value in relation to human ends, or at best intrinsic value that is conditional on their role in human life. The dream of AGI, should it ever become a reality, will generate an even more acute version of this problem, given the prominent role that our rational capacities play in elevating human dignity above the dignity of other beings known to us.[29] For the foreseeable future, however, our focus must be on properly integrating AI technology into a culture that respects and advances the dignity and well-being of humans, and the nonhuman animals with whom we share the world, rather than on the highly speculative endeavor of integrating the dignity of intelligent machines into our existing ethical framework.

AUTHOR'S NOTE

This essay began life as a blog post for the Ada Lovelace Institute, "The Role of the Arts and Humanities in Thinking about Artificial Intelligence (AI)." I am grateful to the Institute for permitting me to reuse some material here. I have benefited from comments on previous drafts from Dominic Burbidge, Hélène Landemore, Seth Lazar, Ted Lechterman, James Manyika, Adrian Vermuele, Carina Prunkl, Divya

Siddarth, Carissa Veliz, Glen Weyl, Mike Woolridge, and John Zerilli. I regret that I have not been able to pursue many of their very stimulating comments within the confines of this short essay.

ABOUT THE AUTHOR

John Tasioulas is Professor of Ethics and Legal Philosophy at the Faculty of Philosophy and Director of the Institute for Ethics in AI at the University of Oxford. He is the editor of *The Cambridge Companion to the Philosophy of Law* (2020) and *The Philosophy of International Law* (with Samantha Besson, 2010).

ENDNOTES

[1] I shall assume a very broad understanding of AI as essentially the use of machines to perform tasks that characteristically require intelligence when performed by humans. My focus will primarily be on "narrow" AI applications, such as facial recognition, surveillance, and risk-assessment, rather than artificial general intelligence, though I say something about the latter toward the very end.

[2] Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (New Haven and London: Yale University Press, 2021), 224.

[3] Some of these issues are compellingly developed by Joshua Cohen in "Don't Shoot the Algorithm" (unpublished manuscript).

[4] The "effective altruism" movement, which has significant allegiance among tech elites, is arguably one expression of this depoliticized and, in its effect, ultimately conservative view of ethics. See Amia Srinivasan, "Stop the Robot Apocalypse," *London Review of Books* 37 (18) (2015).

[5] John Rawls, *Political Liberalism* (New York: Columbia University Press, 1993). For an attempt to pursue the radical discontinuity thesis in relation to AI, see Iason Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds and Machines* 30 (3) (2020): 411.

[6] See John Tasioulas, "The Liberalism of Love," in *Political Emotions: Towards a Decent Public Sphere*, ed. Thom Brooks (London: Palgrave Macmillan, forthcoming 2022).

[7] I am here identifying an influential mode of thought that Russell's book epitomizes. It should be emphasized, however, that there have always been scientists in this domain who have urged the importance of a multidisciplinary approach with an important humanistic dimension, such as in Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (San Francisco: Freeman & Co, 1976); and, more recently, in Nigel Shadbolt and Roger Hampson, *The Digital Ape: How to Live (in Peace) with Smart Machines* (London: Scribe, 2018).

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

[9] Ibid., 255.

[10] Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein, *Noise: A Flaw in Human Judgment* (London: William Collins, 2021).

[11] Ibid., chap. 10.

[12] For a discussion of studies showing that AI predictive tools made no real difference in diagnosing and triaging COVID-19 patients, and in some cases, may have been harmful, see William Douglas Heaven, "Hundreds of AI Tools Have Been Built to Catch Covid, None of Them Helped," *MIT Technology Review*, July 30, 2021, https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/.

[13] John Tasioulas, "The Rule of Law," in *The Cambridge Companion to the Philosophy of Law*, ed. John Tasioulas (Cambridge: Cambridge University Press, 2020), 131–133.

[14] Joseph Raz, *The Morality of Freedom* (Oxford: Oxford University Press, 1986), chap. 12.

[15] Anca Gheaus and Lisa Herzog, "The Goods of Work (Other Than Money!)," *Journal of Social Philosophy* 47 (1) (2016): 70–89.

[16] 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-for.

[17] For an exploration of this theme, see Frank Pasquale, *New Laws of Robotics: Defending Human Expertise in the Age of AI* (Cambridge, Mass.: Harvard University Press, 2020).

[18] For a powerful recent defense of democracy along these lines, see Josiah Ober, *Demopolis: Democracy Before Liberalism in Theory and Practice* (Cambridge: Cambridge University Press, 2017).

[19] For a candid statement, by a Silicon Valley billionaire, of the need to harness the libertarian promise of technology to an antidemocratic ethos, see Peter Thiel, "The Education of a Libertarian," *Cato Unbound: A Journal of Debate*, April 13, 2009, https://www.cato-unbound.org/2009/04/13/peter-thiel/education-libertarian.

[20] For a helpfully wide-ranging discussion of some of these issues, see Joshua Cohen and Archon Fung, "Democracy and the Digital Public Sphere," in *Digital Technology and Democratic Theory*, ed. Lucy Bernholz, Hélène Landemore, and Rob Reich (Chicago: University of Chicago Press, 2021).

[21] For some positive thinking along these lines, see Hélène Landemore, "Open Democracy and Digital Technologies," in *Digital Technology and Democratic Theory*. For useful discussions of digitally enhanced democracy in pioneering countries such as Estonia and Taiwan, see Hans Kundani, *The Future of Democracy in Europe: Technology and the Evolution of Representation* (London: Chatham House, 2020), https://www.chathamhouse.org/sites/default/files/CHHJ7131-Democracy-Technology-RP-INTS-200228.pdf; and Divya Siddarth, *Taiwan: Grassroots Digital Democracy That Works* (New York: Radical Exchange, 2021), https://www.radicalxchange.org/media/papers/Taiwan_Grassroots_Digital_Democracy_That_Works_V1_DIGITAL_.pdf.

[22] "More Than Half of Europeans Want to Replace Lawmakers with AI, Study Finds," CNBC, May 27, 2021, https://www.cnbc.com/2021/05/27/europeans-want-to-replace-lawmakers-with-ai.html. For an interesting discussion of "algocracy" ("rule by algorithm"), see Ted Lechterman, "Will AI Make Democracy Obsolete?" *Public Ethics*, August 4, 2021, https://www.publicethics.org/post/will-ai-make-democracy-obsolete. It is worth noting that proposals for algocracy often assume that the point of politics is

to aggregate human preferences, in line with the preference-based utilitarianism discussed above.

[23] For a perceptive discussion of the way AI threatens to disrupt the economic underpinnings of democracy, see Daron Acemoglu, *Redesigning AI: Work, Democracy, and Justice in the Age of Automation* (Boston: Boston Review Forum, 2021).

[24] Like many others, Stuart Russell adopts an impoverished conception of intelligence as competence in means-ends reasoning according to which the choice of ends is made extraneous to the operations of intelligence. On this view, a machine that annihilated humanity in order to maximize the number of paper clips in existence can qualify as superintelligent. Ibid., 167. For a wide-ranging and perceptive discussion of problems with the idea of "intelligence" invoked in discussions of AGI, see Divya Siddarth, Daron Acemoglu, Danielle Allen, et al., "How AI Fails Us" (Cambridge, Mass.: Edmond J. Safra Center for Ethics, 2021).

[25] Some of these concerns are discussed in John Tasioulas, "First Steps Towards an Ethics of Robots and Artificial Intelligence," *Journal of Practical Ethics* 7 (1) (2019): 61–95, http://www.jpe.ox.ac.uk/papers/first-steps-towards-an-ethics-of-robots-and-artificial-intelligence/.

[26] Kazuo Ishiguro, *Klara and the Sun* (London: Faber, 2021), 218, 306.

[27] Nicholas Wolterstorff, *Justice: Rights and Wrongs* (Princeton, N.J.: Princeton University Press, 2008), 352–361.

[28] David Wiggins, *Solidarity and the Root of the Ethical* (Lawrence: The Lindley Lecture, University of Kansas, 2008).

[29] See, for example, John Tasioulas, "Human Dignity and the Foundations of Human Rights" in *Understanding Human Dignity*, ed. Christopher McCrudden (Oxford: Oxford University Press, 2013), 293–314; and Jeremy Waldron, *One Another's Equals: The Basis of Human Equality* (Cambridge, Mass.: Harvard University Press, 2017).