doi:10.7551/mitpress/14157.001.0001

7 SYSTEMS JUSTICE, AI, AND THE MORAL IMAGINATION

Vafa Ghazavi

Introduction: The Inescapable Present

We live in a deeply unjust world. Across borders, life chances are hugely affected by where we are born. The median age for someone in Uganda or Niger is below sixteen years while in Germany or Japan it is around forty-six.[1] One in four children suffer stunted growth, rising to one in three in developing countries.[2] Around four billion people lack any access to the internet.[3] Within countries, social, gender, and racial disadvantages limit, often sharply, prospects for a flourishing life. Frequently such disadvantages are deeply intertwined with pervasive patterns of social and economic life. As recent research by Raj Chetty and his collaborators has shown, for instance, black boys from wealthy families in America are more likely to become poor in adulthood compared to their similarly wealthy white peers. Even when they grow up in the same neighborhood with parents at the same income level, black boys end up with lower incomes than white boys in 99 percent of the country.[4] Arbitrary limitations on "development as freedom," in Amartya Sen's famous formulation, abound.[5] And beyond inequality and oppression, though not disconnected from it, lies humanity's reckoning with climate change

and what some have described as the Anthropocene. The latter concept captures the profound intensification of human impact on the environment in our age, which, according to Jedediah Purdy, "finds its most radical expression in our acknowledgment that the familiar divide between people and the natural world is no longer useful or accurate."[6]

Against this backdrop, prevailing conceptions of responsibility for realizing justice are under immense strain. In our world, causal connections to unjust harms are often highly diffuse, and the effects of discrete actions on overall outcomes are increasingly hard to discern. Even as connections between deprivations and globalized social processes have intensified, moral implications for specific agents remain unclear. The standard view of "normal justice," as the political theorist Judith Shklar called it, comes to represent the interests of some, often a relatively privileged few, while neglecting those of others.[7] Meanwhile, few alternatives to dominant market or political dogmas appear to exist that can adequately respond to these circumstances. Individual virtue seems insufficient, but institutional responses to take up the slack are not in sight either.

This is the inescapable context in which artificial intelligence is surging as a global political, economic, and cultural force. Yet this backdrop rarely makes it into the discussion of how machines ought to fit into our collective future. We cannot succumb to technological path dependencies. A defining question of our age is whether machines will exacerbate current injustices or contribute to rooting them out.

Reductionism and the Limits of Moral Mathematics

Reductionism has a firm grip on contemporary thinking about justice. At least since the influence of Descartes and Newton

on scientific method in the seventeenth century, reductionism has been a powerful force on all aspects of Western thought.[8] It has also been a check on moral and political imagination. This generates a blind spot for harms that emerge from many people and organizations—most prominently corporations and states—pursuing their goals and interests within the limits of accepted rules and norms. That is, there is a blind spot in ethics for harms that emerge from interconnected, complex systems. Liberal theories of justice are generally inadequate to solve this problem. By focusing on establishing institutions to remediate and mitigate unjust outcomes without also promoting processes of social transformation needed to prevent those injustices, they narrow imagination precisely where it needs to be expanded. This is a major deficiency since—within the status quo, baseline morality of the age—either the injustices in question are rendered invisible or correlative duties to rectify them cannot be located. In other words, no one appears to be responsible for systemic harms. After all, individual components of the system are not necessarily breaching a threshold of blameworthiness, and this is what counts in much contemporary Anglo-American philosophy.

To see what is at stake, contrast the reductionist view with the vision of responsibility articulated by Simone Weil, the great mystic and intellectual, in her 1949 masterpiece on the future of France, *The Need for Roots*:

> Initiative and responsibility, to feel one is useful and even indispensable, are vital needs of the human soul . . .
>
> For this need to be satisfied it is necessary that a man should often have to take decisions in matters great or small affecting interests that are distinct from his own, but in regard to which he feels a personal concern. . . . He requires to be able to encompass in thought the entire range of activity of the social

> organism to which he belongs. . . . For that, he must be made acquainted with it, be asked to interest himself in it, be brought to feel its value, its utility and, where necessary, its greatness, and be made fully aware of the part he plays in it.
>
> Every social organism, of whatever kind it may be, which does not provide its members with these satisfactions, is diseased and must be restored to health.[9]

The liberal orthodoxy in contemporary philosophy, to say nothing of economics, does not promote this rich conception of human responsibility. It implicitly views the individual as an atomized, self-interested agent for whom moral obligations are a burden to be avoided unless and until there is a recognized violation of justice or harm.

This paradigm generates distinctive future risks. For technologists, investors, and geopolitical strategists alike, ideas of responsibility tend to be tied to notions of linear causality or narrow fiduciary roles such as the firm's responsibility to its shareholders. This approach weighs moral costs and duties as if it were an accounting ledger. It focuses on interactions between agents or the rules governing such interactions but pays little attention to system goals or paradigms. In the context of profit-driven or geopolitical competition, individual advances in AI can lead cumulatively to unanticipated harmful outcomes. Such competition is likely to intensify as the gains from AI become more central to economic and social outcomes.

The grip of this competitive dynamic threatens to co-opt, marginalize, or crowd out efforts to orient AI development toward the common good. As Stephen Cave and Seán Ó hÉigeartaigh point out, framing AI development as a "race for technological superiority" can create serious societal risks. Even if such a competitive race is not actually pursued, they

argue, its narrative can generate a politics of fear or insecurity that erodes trust, limits deliberation, and dampens collaboration on an AI agenda that promotes collective benefits. Moreover, such rhetoric can itself spark a race for technological advantage.[10] Perhaps most troubling, I think, is the possibility that different actors begin to feel that such a trajectory is inevitable and adjust their moral baselines in light of that perceived reality. Once such a mindset takes hold, the prospect of pursuing an alternative paradigm diminishes. The race dynamic could become self-sustaining. What are the alternative narratives that could be promoted? Among other strategies, Cave and Ó hÉigeartaigh suggest reframing AI development as a shared priority for global good, including by emphasizing its potential to tackle large-scale challenges such as climate change and poverty. Shifting the emphasis in this way, in their view, could help counteract a race dynamic by downplaying the importance of which companies or countries make key breakthroughs, highlighting the mutual benefits of cooperation in the face of global challenges, and including the global community as stakeholders in the process of AI development.

The pressing task for our collective future with machines is therefore not simply to predict the risks, however important that might be. Nor is it to usher in a utopian Singularity. Rather, it is to imagine and then continually make and remake a world where scientific discovery and emergent technologies deepen human flourishing. This involves discerning what we value most, individually and collectively, rather than simply adjudicating ethical dilemmas or structuring society to compensate for inequalities or harms *after* the market has produced them. Even if we accept that market forces can drive productive innovation in AI, we have to ask whether the incentives embedded in the market are themselves congruent with wider social

and moral purposes. Market competition left to itself cannot supply this. It is hard to resist the force of John Palfrey's argument that "it feels urgent that we examine what we care most about in humanity as we race to develop the science and technology of automation" (or, we can add, AI more broadly).[11] An ethical framework based on backward-looking blame and guilt won't be up to the great task before humanity to "design systems that participate as responsible, aware, and robust elements of even more complex systems."[12]

The dominant view of harm, even in large-scale, highly diffuse cases of systemic injustice, is one of assigning blame for discrete wrongdoing. But this model of carving up responsibility struggles when an agent's marginal contribution to a particular harm is almost unidentifiable. Moral philosophers such as Derek Parfit—a towering figure in contemporary analytic philosophy—have persuasively challenged our intuitions on this sort of problem.[13] Their arguments expose pitfalls of our "moral mathematics" and suggest that culpability of some kind is called for even when actions appear to fall below a threshold for wrongdoing. But they do little to overcome the difficulty of distinctly *systemic* harm. Even in its aggregated form, the reductionist account depends on a linear model of causality and rectification of harm, with agents connected to outcomes in predictable ways. The default becomes to blame a few exceptional wrongdoers, such as those most clearly linked to the endpoint of harm, rather than to see the wholeness of the situation.

This is unsurprising given the grip of salience and the availability heuristic on human psychology.[14] But as this type of connection diminishes or dissolves entirely through system effects that cannot be foreseen at the outset by the doers or enablers of harm, the limitations of the reductionist account

become more apparent. What we need are normative reasons to creatively transform the existing incentive structure or its compliance regime, not simply mechanisms to allocate responsibility more precisely.[15] Rather than satisfying itself by closing in on a narrow set of moral duties, our theory of justice should liberate the very imaginative context that dictates the limits of harm, ethics, and virtue.

A more useful way to address this challenge is the concept of *structural injustice* developed by the political theorist Iris Marion Young before her untimely death in 2006. Structural injustice consists of social processes that "put large groups of persons under systematic threat of domination or deprivation of the means to develop and exercise their capacities, at the same time that these processes enable others to dominate or to have a wide range of opportunities for developing and exercising capacities available to them."[16] In response, Young advocated a "social connection model" of responsibility, in contrast to a legalistic liability model. According to this, "all those who contribute by their actions to structural processes with some unjust outcomes share responsibility for the injustice."[17] Young argued that liability, deriving from notions of guilt or blame for wrongdoing, is an inappropriate framework for assigning responsibility in relation to structural injustice. Liability fixates on guilt, which is unhelpful because it directs attention to some actors while absolving others, deflects attention from background conditions, and produces defensiveness, creating division where unified action is called for.

Instead, Young's model promotes *political responsibility*, which consists of an imperative to watch social institutions, monitor their effects "to make sure that they are not grossly harmful," and maintain "organized public spaces where such watching and monitoring can occur and citizens can speak

publicly and support one another in their efforts to prevent suffering."[18] The meaning of politics here is "public communicative engagement with others for the sake of organizing our relationships and coordinating our actions more justly."[19]

To better grasp what this looks like in practice, consider the social transformation that has been taking place in response to the problem of modern slavery and labor exploitation in global supply chains. As Young pointed out, labor exploitation in supply chains was not widely considered a problem of the multinational firms or their consumers in the rich world until fairly recently. Rather, it was posed as a problem of unethical behavior or lax regulation/enforcement in the poorer countries where the exploitation took place. It was a moral problem for the factory manager in Bangladesh, not the CEO or consumer in New York. It took a qualitative shift in imagination supported by social movements and ordinary consumers to bring all the connected agents into the same moral frame. Indeed, the approach moved from one of interactive ethics—the ethics of each interaction in the market—to one of judging the morality of how agents are positioned within wider market and social dynamics. How we collectively view labor exploitation in global supply chains is gradually moving from a reductionist logic to one of systems, with all the political and economic implications entailed by such a move.

In thinking about the responsibility of specific agents to address structural injustices, Young argues that we ought to weigh several parameters, including an agent's power, privilege, interests (not least those of victims themselves), and ability to work collectively with others. I suggest something analogous is needed for developments in AI. To link the development of machines to human flourishing, a wide array of individual and collective agents will need to take up responsibilities to monitor

system effects, call out destructive dynamics, and help model and construct alternative practices and paradigms.

It is striking that contemporary political philosophers have paid so little attention to the normative implications of complex systems. This vantage point connects agency with systemic change in a morally and psychologically compelling way with significant implications for the public sphere. Much of our thinking about justice has gone astray because it has failed to account for system effects in either the perpetuation of injustice or the realization of justice over time. Shifting the emphasis to system goals and paradigms, however, brings moral and political imagination center stage. This opens quite radical departures from prevalent thinking with regard to *who* has responsibilities for justice, *how* these responsibilities are fulfilled, and *what* the very nature of such responsibilities is. Taken together, this encourages me to propose a new way of thinking about justice, which I call *systems justice*.

Systems Justice

At the height of the global financial crisis in November 2008, on a visit to the London School of Economics, Her Majesty Queen Elizabeth II of England asked a group of assembled academics a succinct but piercing question: "Why did no one see it coming?"[20] In June of the following year, a group of leading experts met at the British Academy to try and answer the Queen's question. Their response, distilled in a letter, included this: "So in summary, Your Majesty, the failure to foresee the timing, extent and severity of the crisis and to head it off, while it had many causes, was principally a failure of the collective imagination of many bright people, both in this country and internationally, to understand the risks to the system as

a whole."[21] Such failures of imagination afflict not only technocrats and economists but also the wide range of agents in complex social systems. The inability to appreciate the "system as a whole" has deep intellectual and cognitive roots. People experience systems differently to interactions with other agents, rules, and material facts. James Scott, for instance, has documented the disastrous consequences when states seduced by "high modernism" have established schemes that render complex social realities legible to control according to scientific laws.[22] More generally, humans tend to demonstrate a low aptitude for learning from complex interactions, focusing on end-state *outcomes* rather than what might have been, the *counterfactual*; we rationalize how we got to where we are as a natural and inevitable product of past events.[23] It is unsurprising, then, that we reify and naturalize existing systems.

Recognizing the ubiquity of system effects as the starting point for thinking about justice on a global scale, however, is a way out of this reductionist trap. It generates a new vantage point that unlocks the potential of each person to play their part in imagining, experimenting in, and realizing a more just world. To envision the possibilities, we can turn to the great historical struggles against slavery, racism, and patriarchy.[24] Faced with structural injustice, the response of many agents throughout history has been to seek to undermine the future durability of the system of oppression—especially its discourse and sentiments—in ways that cannot be reduced to a simple formula of obligations. Such responses do not necessarily resemble the repairing of harm to specific victims but rather look like an attack on background injustice through an expansion of collective moral imagination. In this task, relying on a precise, measurable division of responsibility within the status quo morality of formal rules and recognized institutions may

even inhibit our capacity to notice unjust system-level harm. Indeed, discharging such duties can promote the illusion that we have done our fair share.

Responsibility for justice is thus tied to our *sense of self* as moral beings: there is ultimately no meaningful distinction here, I suggest, between the interests and life projects of individual agents and affirmative responsibility for justice. This suggests the paradigm should therefore be one of *coherence*, rather than dividing the costs of action. This requires citizens—including technologists, investors, and public leaders of various kinds—committed to human flourishing in a larger sense, beyond what is formally demanded of them. It is about virtue and integrity. Our theory of justice, if it is to seriously grapple with the complexity of harm in a globalized world, must make room for this constructive view of responsibility. It should encourage people to identify how they can positively apply their talents, life projects, and social roles to the transformation of unjust systems. Such a project must be reflected in the ethos of educational programs so as to enable all citizens to develop their capabilities to contribute to the process of structural transformation.

This task—one of empowerment and ennoblement—was already urgent given the challenges of global injustice and impending ecological crisis, but the advancement of AI accelerates this urgency further still. The interests driving advances in AI can either join this agenda or ignore it. But if they choose the latter, the cumulative effect will likely end up moving closer to and eventually coinciding with a resounding answer to Kwame Anthony Appiah's question: "What will future generations condemn us for?"[25]

To see what is already at stake, consider Cathy O'Neil's work on "weapons of math destruction" (or WMDs)—her

descriptor for harmful mathematical models underpinning algorithms that power the data economy. As O'Neil, a data scientist and writer, acknowledges, WMDs reflect the "choices made by fallible human beings." Even when made with the best intentions, many of these models, she observes, encode "human prejudice, misunderstanding, and bias" into software systems that are now such a powerful force in everyday life.[26] What is particularly important for my argument is the way in which incentives embedded in the social system can drive these harmful outcomes through individual actions that are still within the bounds of moral acceptability. O'Neil points out the feedback and incentive for those running these models is profit: "Their systems are engineered to gobble up more data and fine-tune their analytics so that more money will pour in. Investors, of course, feast on these returns and shower WMD companies with more money."[27] Those involved in such processes invariably push for whatever marginal gains they can—the next technological advance to make that extra bit of profit—detached from the structural effects that may have a bearing on justice or human flourishing more broadly. There is a double moral blind spot here: one in terms of the immediate effects of an algorithm on identifiable people, and the other in terms of reinforcing the orientation of the system as a whole.

What is needed in response is a wider conception of responsibility that includes reenvisioning the underlying systems that allow such algorithms to proliferate without regard to their social impact. But what could ground this idea of responsibility? By acknowledging the potential of all of us in our situated social roles—as, for example, technologists, artists, intellectuals, politicians, CEOs, investors, parents, citizens—to contribute to social transformation, the project of systems justice deepens our sense of who we want to become as moral beings.

It is a provocation and invitation to do our part—humbly and with openness—in an ongoing, dynamic process of change. This approach pushes back against the premise of an atomized, self-interested agent and promotes alternative grounds for commitments to the common good.

The prevailing cost- or liability-based model suggests that some agents must be held responsible for a harm so that there is not an undue imposition on the lives of ordinary people going about their lives in morally legitimate ways. But this division is precisely what is at stake. The ends we seek must themselves be conceptualized within a given social condition in which we find ourselves. It is a moral mistake, I think, to suggest that we can partition ourselves from the world, demarcating one realm as that of personal freedom and autonomy and the other as that of social relations or the natural environment. In truth, the two are bound together. Fundamentally, systems justice starts by asking who we are and want to become as moral beings, rather than asking what costs we owe or burdens we should take on.[28] Since virtue is fortified in response to injustice, our response to injustice can itself become an aspect of the good life and human flourishing. Human interests therefore should not be viewed as fixed. As we begin to engage in system transformation, our interests and those of others can transform. After all, a flourishing life depends at least in part on our sense of something bigger than ourselves.[29]

Systems justice thus connects a bird's-eye view of justice—consistent with the "impartial spectator" deployed in Amartya Sen's conception of justice[30]—to the distinctive position of each agent in a social system. It encourages us to see the wholeness of the situation and to design institutions in light of the "admissibility of incompleteness."[31] The focus shifts from a

perfectly just society derived from just institutions to a comparatively just society focused on social realizations.[32]

Systems justice, then, is neither an analytical category nor the articulation of a particular state of affairs. Rather, it is a lens through which moral agents can see the world from different vantage points and motivate their distinctive contribution to meet the moral needs of the age, including through deepening the *ethos* of justice—a concept the philosopher G. A. Cohen described as "a structure of response lodged in the motivations that inform everyday life."[33] It is a way to interpret our responsibilities given the exigencies of the multiple, nested social systems in which we live, not a formula for the design of perfect institutions. It is an ontological move that draws on complex systems as a recurring metaphor for imagining social and political life and the relationships agents have to each other over time. It is a standpoint for discerning individual system-level obligations of justice. Most importantly, it generates a practical *morality of coherence*—one that connects small-scale action and commitments with large-scale transformation—over a *compartmentalized morality*. Reading the essay by Lewis, Arista, Pechawis, and Kite (chapter 1), I am struck by how this approach mirrors Indigenous epistemologies that emphasize relationality.[34] I find this reassuring since it suggests that my theory may have particular purchase in confronting the challenges and opportunities of AI if we orient ourselves in a certain direction toward this task.

To be clear, my point is not to suggest that the direct responsibility of technologists or regulators, for example, is unimportant. Rather, it is to say that *both* direct and system responsibility matter. We should think of liability and social connection as *complementary* elements of a far-reaching vision of responsibility. Accountability can be effectively linked to systemic change.

Incentive-based measures to promote direct liability, however, must be sensitive to inadvertently shifting the logic from one of morality to one of maximizing gains, thus crowding out moral motivations.[35] Crucially, the manner and language through which perpetrators of injustice are held accountable should enrich and sharpen, rather than diminish, our sense of shared responsibility. In my approach, citizens become coauthors of justice rather than passive recipients of arrangements enacted and enforced from the top down. Analogously, holding powerful technology companies accountable is a call to wider moral agency, not a means of letting others off the hook.

It seems we do not have well-developed conceptual resources for the ambiguous space—the transitional phase—between injustice and justice. But it is in this in-between space that real-world contestation over moral ideas and their practical expression takes place. Incremental contributions are often an essential feature for these transitions.

Experimentalism and Hope

The moral implications of AI therefore require a new kind of responsibility. Agents—citizens, firms, states—must weigh their ethical responsibilities in relation to their connection to the entire system, not only discrete, linear interactions or even their simple aggregation. Systems justice subverts how we think about who has moral responsibility and how it is fulfilled. Since discourses and sentiments of the system shape it more than its mechanics, the role of artists and writers, for example, comes to the fore; small acts of moral courage become more powerful than they seem in our conventional ethical calculus.

Systems justice thus resists hasty conclusions on negligibility, the suggestion that small-scale contributions are largely

irrelevant to structural outcomes. Instead, it highlights the importance of ordinary attitudes and behaviors in sustaining—or transforming—social systems. As the Plato scholar Melissa Lane suggests in her penetrating book that draws on ancient ethics to respond to contemporary challenges of environmental sustainability: "The person who embodies a new outlook becomes in virtue of that very fact a node in a new political imagination, the first step to creating a new social ethos."[36] Indeed, systems justice incorporates a moral version of the concept of the "butterfly effect" made famous by Edward Lorenz. As the feminist legal scholar Catherine MacKinnon more recently described this idea, "some extremely small simple actions, properly targeted, can come to have highly complex and large effects in certain contexts."[37] This yields a deceptively simple normative insight: the power for large-scale transformation is already latent within the system. "If no paradigm is right," Donella Meadows writes, "you can choose whatever one will help to achieve your purpose."[38]

While social scientists have sought to explain society-wide normative change through concepts such as norm, reputational, and availability *cascades* and *tipping points*,[39] these empirical phenomena have been neglected in the formulation of normative theory itself. Looking back to examples of social transformation does not yield fine-grained answers on what moral agents are obliged to do in the face of structural injustice. Rather, it suggests the wide scope moral agents have to reconceive their interests and moral identity, and the diverse ways they can direct their lives toward the realization of justice.[40] These studies reveal that the transformation of discursive and imaginative context lies within the grasp of moral agents, especially when acting in concert. They bring to light, for example, how the force of example and the generation of inertial momentum through incremental steps can reshape the

social ethos and disrupt destructive path dependencies. This is most evident among an avant-garde who promote a new standard of justice,[41] but a wide array of agents, including ordinary citizens, play an indispensable role. Some of these contributions may not be salient or even visible, except perhaps retrospectively, yet this does not diminish their moral force.

What does this mean for the development of AI? At least one implication pertains to overcoming "the myth of people as socially independent," which, as Molly McCue and Kat Holmes point out, "not only limits who can participate in the system, but also who can contribute to the evolution of that system through design."[42] They pose a crucial question: "If we develop our innate ability to connect with one another as a precious resource and source of social vitality, what kind of AI could we build?"[43]

Since agents cannot predict or control the evolution of social systems ex ante, systems justice includes a normative commitment to democratic experimentalism. This principle draws on the tradition of American pragmatism found in the work of John Dewey and William James.[44] Under systems justice, institutions and citizens alike commit themselves to *discovering* harms and injustices, and innovations to reduce them over time. Systems justice thus denaturalizes current forms of the market and governance, leaving them open to revision. It harnesses people's distinctive talents and capabilities in this enterprise, promotes unorthodox alliances and multi- and "antidisciplinary" innovations,[45] and strengthens mechanisms for social learning about effective ways to reduce structural injustice and widen the development of human capabilities. Jaclyn Sawyer's essay in this volume on how social workers can shed light on history and social context so as to enrich the work of technologists illustrates what this might look like

in practice.[46] Such collaboration between social workers and technologists, Sawyer argues, can help confront those heuristics that lead to a disconnect between the intention of a technology and its impact on human lives. We need much more of this kind of collaboration if AI is going to advance human flourishing.

Conclusion

Our future with machines looks set to be defined by the cumulative, often unintended consequences of many agents working to advance the frontiers of intelligence within existing rules and norms. Traditional models of responsibility are inadequate to confront this reality. The moral challenge in these circumstances is to foster citizens and design institutions that are responsive to these system-level effects. Beyond governments, agents such as individuals, civil society groups, and corporations can play a surprisingly constructive role by drawing on their distinctive talents, knowledge, and capacities. Since these agents cannot predict or control the evolution of social systems ex ante, citizens and institutions must remain open to discovering systemic injustice, innovating to root it out, and reconceiving their own interests in the process. Responsibility for systemic harm is fundamentally shared. Shirking this is not only problematic for potential victims but also corrodes the virtue and moral identity of the irresponsible agent. Ultimately, freedom for the self is connected to promoting justice for the whole since the two are permanently intertwined, part of a single moral life. Resisting reduction opens possibilities not only for extended intelligence but also for extended morality. Obligations of justice in any given moment of history—and the values we embed in the process

of technological evolution—are therefore not impositions or burdens, but rather the means to become who we aspire to be as moral beings.

Notes

For discussions on various themes and stray ideas related to this essay, I am grateful to William Butler, Daniel Butt, Janina Dill, Joi Ito, Cécile Laborde, Amartya Sen, Henry Shue, Kathryn Sikkink, Roberto Mangabeira Unger, James Walsh, Jonathan Zittrain, and, especially, Jonathan Wolff.

1. "Global Health Observatory Data Repository," World Health Organization, http://apps.who.int/gho/data/view.main.POP2040ALL?lang=en.

2. "Goal 2: Zero Hunger," United Nations, https://www.un.org/sustainabledevelopment/hunger/.

3. "ICT Facts and Figures 2017," International Telecommunication Union, https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx.

4. Raj Chetty, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter, "Race and Economic Opportunity in the United States: An Intergenerational Perspective," NBER Working Paper No. 24441, March 2018.

5. Amartya Sen, *Development as Freedom* (New York: Knopf, 1999).

6. Jedediah Purdy, *After Nature: A Politics for the Anthropocene* (Cambridge, MA: Harvard University Press, 2015), 2.

7. Judith N. Shklar, *The Faces of Injustice* (New Haven and London: Yale University Press, 1990).

8. René Descartes, *A Discourse on the Method of Correctly Conducting One's Reason and Seeking Truth in the Sciences* (Oxford: Oxford University Press, [1637] 2006), 17. Descartes summarized his approach in the following way:

> I came to believe that in the place of the great number of precepts that go to make up logic, the following four would be sufficient for my purposes, provided that I took a firm and unshakeable decision never once to depart from them. The first was never to accept anything as true that I did not *incontrovertibly* know to be so; that is to say, carefully to avoid both *prejudice* and premature conclusions; and

> to include nothing in my judgements other than that which presented itself to my mind so *clearly* and *distinctly*, that I would have no occasion to doubt it. The second was to divide all the difficulties under examination into as many parts as possible, and as many as were required to solve them in the best way. The third was to conduct my thoughts in a given order, beginning with the *simplest* and most easily understood objects, and gradually ascending, as it were step by step, to the knowledge of the most *complex*; and *positing* an order even on those which do not have a natural order of precedence. The last was to undertake such complete enumerations and such general surveys that I would be sure to have left nothing out.

9. Simone Weil, *The Need for Roots: Prelude to a Declaration of Duties towards Mankind* (New York: Routledge, [1949] 2002), 15.

10. Stephen Cave and Seán S. Ó hÉigeartaigh, "An AI Race for Strategic Advantage: Rhetoric and Risks," paper presented at the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society, February 5, 2018, http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_163.pdf.

11. John Palfrey, "Line-Drawing Exercises: Autonomy and Automation," *Journal of Design and Science*, no. 3 (2017), https://jods.mitpress.mit.edu/pub/issue3-palfrey.

12. Joichi Ito, "Resisting Reduction: A Manifesto," *Journal of Design and Science*, no. 3 (November 2018), https://jods.mitpress.mit.edu/pub/resisting-reduction.

13. Derek Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984), 78–84.

14. Amos Tversky and Daniel Kahneman, "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology* 5, no. 2 (September 1973): 207–232; Daniel Kahneman, *Thinking, Fast and Slow* (New York: Farrar, Straus and Giroux, 2011), 129–136.

15. See Iris Marion Young, *Responsibility for Justice* (Oxford: Oxford University Press, 2004), 375.

16. Young, *Responsibility for Justice*, 52.

17. Young, *Responsibility for Justice*, 96.

18. Young, *Responsibility for Justice*, 88.

19. Young, *Responsibility for Justice*, 179.

20. Chris Giles, "The Economic Forecasters' Failing Vision," *Financial Times*, The FT Year in Finance supplement, December 16, 2008, 5.

21. Tim Besley and Peter Hennessy, "The Global Financial Crisis—Why Didn't Anybody Notice?," *British Academy Review* 14 (November 2009): 8–10.

22. James C. Scott, *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* (New Haven and London: Yale University Press, 1998).

23. Duncan J. Watts, *Everything Is Obvious: Once You Know the Answer* (New York: Crown Business, 2011).

24. On slavery, see, for example, Manisha Sinha, *The Slave's Cause: A History of Abolition* (New Haven and London: Yale University Press, 2017); Adam Hochschild, *Bury the Chains: The British Struggle to Abolish Slavery* (London: Macmillan, 2005); Neta C. Crawford, *Argument and Change in World Politics: Ethics, Decolonization, and Humanitarian Intervention* (Cambridge: Cambridge University Press, 2002). On women's suffrage (as one illustration of confronting patriarchy), see, for example, Martha Finnemore and Kathryn Sikkink, "International Norm Dynamics and Political Change," *International Organization* 52, no. 4 (Autumn 1998): 887–917; Francisco O. Ramirez, Yasemin Soysal, and Suzanne Shanahan, "The Changing Logic of Political Citizenship: Cross-National Acquisition of Women's Suffrage Rights, 1890 to 1990," *American Sociological Review* 62, no. 5 (October 1997): 735–745. One of the lessons from these historical cases is that the more granular perspective we take, the more it becomes apparent that ordinary citizens have been crucial to transformational change even if less valorized compared to a few exceptional leaders. For example, it might easily be forgotten that around the 1790s, 300,000 English people participated in a sugar boycott to abolish slavery (see Hochschild, *Bury the Chains*, 192–196).

25. Kwame Anthony Appiah, "What Will Future Generations Condemn Us For?," *Washington Post*, September 26, 2010, B01; Kwame Anthony Appiah, *The Honor Code: How Moral Revolutions Happen* (New York and London: W. W. Norton and Company, 2010).

26. Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (London: Allen Lane, 2016), 3.

27. O'Neil, *Weapons of Math Destruction*, 13.

28. On the application of this to the public realm, consider the argument in Benjamin R. Barber, *Strong Democracy: Participatory Politics for a New Age* (Berkeley, Los Angeles, and London: University of California Press, [1984] 2003). For instance, on reductionist conceptions of freedom that pose it in opposition to power, Barber remarks: "Rendering freedom and power in physical terms not only misconstrues them, it produces a conception of political liberty as entirely passive. Freedom is associated with the unperturbedness of the inertial body, with the motionless of the inertial frame itself. It stands in stark opposition to the idea of politics as activity, motion, will, choice, self-determination, and self-realization. . . . The modern liberal appears to regard it [tranquility] as a republican ideal: man at rest, inactive, nonparticipating, isolated, uninterfered with, privatized, and thus free" (36).

29. For one argument that offers this view from the perspective of psychology, see Martin E. P. Seligman, *Flourish: A Visionary New Understanding of Happiness and Well-Being* (New York: Free Press, 2011).

30. Sen himself borrows this from Adam Smith. See Amartya Sen, *The Idea of Justice* (Cambridge, MA: Harvard University Press, 2009), 124–152.

31. Sen, *The Idea of Justice*, 131.

32. Sen, *The Idea of Justice*, 134.

33. G. A. Cohen, *Rescuing Justice and Equality* (Cambridge, MA: Harvard University Press, 2008), 123.

34. Jason Edward Lewis, Noelani Arista, Archer Pechawis, and Suzanne Kite, "Making Kin with the Machines," *Journal of Design and Science*, no. 3 (2018), https://doi.org/10.21428/bfafd97b.

35. See Samuel Bowles, *The Moral Economy: Why Good Incentives Are No Substitute for Good Citizens* (New Haven: Yale University Press, 2016).

36. Melissa Lane, *Eco-Republic: What the Ancients Can Teach Us about Ethics, Virtue, and Sustainable Living* (Princeton and Oxford: Princeton University Press, 2012), 64.

37. Catherine A. MacKinnon, *Butterfly Politics* (Cambridge, MA, and London: Harvard University Press, 2017), 1. MacKinnon draws on this concept to name her book, which compiles interventions made over

forty years. She uses the term "butterfly politics" as "an organizing metaphor and central conceit" for the volume.

38. Donella H. Meadows, *Thinking in Systems: A Primer*, ed. Diana Wright (White River Junction, VT: Chelsea Green Publishing, 2008), 164.

39. Seminal examples include Timur Kuran, *Private Truths, Public Lies: The Social Consequences of Preference Falsification* (Cambridge, MA: Harvard University Press, 1997); Cass R. Sunstein, "Social Roles and Social Norms," *Columbia Law Review* 96, no. 4 (May 1996): 903–968; Timur Kuran and Cass R. Sunstein, "Availability Cascades and Risk Regulation," *Stanford Law Review* 51, no. 4 (April 1999): 683–768.

40. To illustrate this, consider how James Baldwin sought to resolve the dilemma he faced when deciding how to play his part in responding to racial injustice in the United States in the mid-twentieth century. After listing other avenues of black resistance and activism and explaining how he did not feel personal compatibility with them, Baldwin writes: "This was sometimes hard on my morale, but I had to accept, as time wore on, that part of my responsibility—as a witness—was to move as largely and as freely as possible, to write the story, and to get it out." See James Baldwin and Raoul Peck, *I Am Not Your Negro* (New York: Vintage Books, 2017), 31.

41. Lea Ypi, *Global Justice and Avant-Garde Political Agency* (Oxford: Oxford University Press, 2012).

42. Molly McCue and Kat Holmes, "Myth and the Making of AI," *Journal of Design and Science*, no. 3 (2018), https://jods.mitpress.mit.edu/pub/holmes-mccue.

43. McCue and Holmes, "Myth and the Making of AI."

44. See, for example, John Dewey, *The Public and Its Problems: An Essay in Political Inquiry* (University Park: The Pennsylvania State University Press, [1927] 2012); William James, *The Will to Believe and Other Essays in Popular Philosophy* (Cambridge, MA, and London: Harvard University Press, [1897] 1979). For more contemporary thinking related to this approach, see Christopher K. Ansell, *Pragmatist Democracy: Evolutionary Learning as Public Philosophy* (Oxford and New York: Oxford University Press, 2011); Michael C. Dorf and Charles F. Sabel, "A Constitution of

Democratic Experimentalism," *Columbia Law Review* 98, no. 2 (March 1998): 267–473; Charles F. Sabel and Jonathan Zeitlin, "Experimentalist Governance," in *The Oxford Handbook of Governance*, ed. David Levi-Faur (Oxford: Oxford University Press, 2012), 169–184.

45. On the idea of an antidisciplinary research program, see Joichi Ito, "Design and Science," *Journal of Design and Science*, no. 1 (2017), https://jods.mitpress.mit.edu/pub/designandscience.

46. Jaclyn Sawyer, "What Social Work Got Right and Why It's Needed for Our (Technology) Evolution," *Journal of Design and Science*, no. 3 (2018), https://jods.mitpress.mit.edu/pub/sawyer.

Against Reduction

Designing a Human Future with Machines

By: Noelani Arista, Sasha Costanza-Chock, Vafa Ghazavi, Suzanne Kite, Cathryn Klusmeier, Jason Edward Lewis, Archer Pechawis, Jaclyn Sawyer, Gary Zhexi Zhang, Snoweria Zhang

Citation:
Against Reduction: Designing a Human Future with Machines
By: Noelani Arista, Sasha Costanza-Chock, Vafa Ghazavi, Suzanne Kite, Cathryn Klusmeier, Jason Edward Lewis, Archer Pechawis, Jaclyn Sawyer, Gary Zhexi Zhang, Snoweria Zhang
DOI: 10.7551/mitpress/14157.001.0001
ISBN (electronic): 9780262367318
Publisher: The MIT Press
Published: 2021

The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Minion and Neue Haas Grotesk by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Arista, Noelani, author.
Title: Against reduction : designing a human future with machines / Noelani Arista, Sasha Costanza-Chock, Vafa Ghazavi, Suzanne Kite, Cathryn Klusmeier, Jason Edward Lewis, Archer Pechawis, Jaclyn Sawyer, Gary Zhexi Zhang, Snoweria Zhang ; introduction by Kate Darling.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Includes bibliographical references and index.
Identifiers: LCCN 2020053013 | ISBN 9780262543125 (paperback)
Subjects: LCSH: Artificial intelligence--Moral and ethical aspects. | Artificial intelligence--Social aspects.
Classification: LCC Q334.7 .A75 2021 | DDC 303.48/34--dc23
LC record available at https://lccn.loc.gov/2020053013