"Justice and Injustice"

Prepared for the Princeton Dialogues on AI & Ethics *Glossary of Technical Terms*

Lucia M. Rafanelli

Introduction

That technical innovation will deliver us not only from productive inefficiency, but also from injustice is a tempting thought. The idea that technology holds within it the seeds of emancipation stretches as far back as Aristotle, who wrote that the invention of autonomous machines would render class hierarchy and slavery obsolete: "...if every instrument could accomplish its own work, obeying or anticipating the will of others...if...the shuttle would weave and the plectrum touch the lyre without a hand to guide them, chief workmen would not want servants, nor masters slaves" (*Politics* 1253b33-1254). Now, in the age of burgeoning AI, this sentiment seems to offer great solace indeed.

But the solace is illusory. AI systems are, ultimately, yet more tools with which people can exercise power over each other. We can exercise our power for good or evil, to promote emancipation or subjugation; we can exercise it accountably or unaccountably, justly or unjustly. So it has always been, and the introduction of AI does nothing to change this.

This entry won't settle the question of what justice requires or of what specific political and social arrangements constitute injustice. It will try to clarify what political theorists and philosophers are talking about when they talk about justice and will highlight some ways in which the use of AI can implicate questions of justice.

What Justice Is—And Isn't

In general terms, "justice" provides *a set of standards by which to fairly adjudicate people's (often competing) claims to various liberties, opportunities, resources, and modes of treatment*. John Rawls, writing about "justice" as applied to a society's major social and political institutions, says that even people who disagree about what justice requires can agree they need *some* standards to serve that function—to determine what people's rights and duties are and how the benefits and burdens of living together in their society should be distributed (Rawls 1999, 5).

Not all political theorists would define justice in Rawlsian terms, however. Some (e.g., Cohen 1997) would challenge Rawls' assumption that justice applies primarily to institutions, arguing that it applies equally to individuals' everyday choices. Others (e.g., Okin 1989) would emphasize that justice doesn't apply only to formal institutions, like the bodies of government, but also to informal institutions, like the family. Still others (e.g., Beitz 1999; Caney 2005; Ypi 2012 and 2013) would argue that principles of justice like those Rawls envisions governing a single society actually apply to everyone around the globe. Nonetheless, we could interpret many who disagree about the precise scope or site of the requirements of justice as endorsing the more

general idea above—that "justice" provides a set of standards by which to fairly adjudicate certain kinds of claims.

A few other ideas about justice are also widely endorsed by political theorists, despite their disagreements about its specific requirements. First, justice is typically thought to be distinct from wellbeing. That a proposed policy would make (some) people richer or happier, for example, doesn't mean it would be just; that it would make (some) people poorer or less happy doesn't mean it would be unjust. Presumably, racist shopkeepers in the American Jim Crow South were displeased when they were legally required to integrate their businesses. Their displeasure, though, did nothing to diminish the fact that justice required integration.

That said, some do think justice and wellbeing are connected, in that people have justice-based claims to a certain level of wellbeing. For example, Henry Shue (1996) argues that people have rights to the goods necessary for subsistence. David Miller (2007, 178-85, 207-8) argues that people have rights to the goods necessary for living a "minimally decent human life." Others (see, e.g., Anderson 1999; Nussbaum 2000 and 2003; Sen 1980) argue that justice requires people to have (or at least have the opportunity to develop) certain capabilities, such as the capability to participate on equal terms with one's fellow citizens in democratic deliberation.

Second, justice is not the same as legality. But, it's often (though not universally) thought that if something is a requirement of justice, it should ideally be a legal requirement, too. Conversely, some argue that the limits on what behaviors we can feasibly guarantee via legal institutions should limit what behaviors we identify as requirements of justice. (Onora O'Neill (2005) makes an argument of this kind cautioning against the classification of what she sees as overly-ambitious guarantees to high-quality healthcare as "human rights.")

Third, there is significant disagreement about what justice requires. But this does *not* mean there is no truth of the matter. People can and frequently do disagree about matters of fact. If you wanted to convincingly argue there was no truth about what justice required, you'd have to do more than show that people disagreed about what that truth was. Nonetheless, taking seriously the idea that people have equal moral worth, and are (therefore) equally entitled to help decide how their political lives are organized, means recognizing the weight of their ideas about what justice requires, even if we believe them to be mistaken. This, in turn, may mean limiting the ways in which we exercise our own political power—for example by helping establish and maintain political institutions that govern in response to their constituents' input, rather than unilaterally imposing our own views of justice on others.

All this is to say that even political theorists who disagree sharply about justice often agree that standards of justice (whatever the right ones are) serve an important social function: they allow us to fairly adjudicate certain kinds of claims. They provide moral, as distinct from legal, standards that tell us how people deserve to be treated—standards that can't be reduced to commands to make people happier or richer. People will continue to disagree about which standards of justice are the right ones. But our actions and interactions, our social and political institutions, will inevitably reflect some ideas about justice at the expense of others. We are continually faced with the question of which of these ideas to privilege, and continually

challenged to make and remake our shared practices and institutions in the service of justice. Few questions are as important as this one, and few challenges as urgent.

Selected Questions of Justice

Figuring out what justice requires necessitates answering countless smaller questions. How should material resources be distributed? If justice requires treating people "as equals," what does this mean—should people be guaranteed equal opportunity, equal social status, equal wellbeing? What kind of influence should people have over their political institutions? What freedoms should governments guarantee to their people? These are only a few prominent questions about justice. I can't hope to answer, or even state, them all here. Alternatively, I will highlight a few ideas about justice that are often overlooked or misunderstood in public political discourse, and that are especially relevant for the ethics of using AI.

Institutional Discrimination and Structural Injustice

Some injustices are created by the cumulative force of countless actions, each perhaps insignificant on its own, and none necessarily undertaken with the aim of oppressing anyone. The actors may have coordinated with each other in the sense of participating together in a shared social system—like their state's legal system, or the global economy. But they need not have coordinated with each other in the sense of intentionally collaborating to advance a shared goal. Each may have acted independently, on her own motives and interests, unaware of the identities, motives, interests, or actions of the others. Nonetheless, their actions, taken together, can produce injustice.

One example is what we might call "institutional discrimination." Tommie Shelby (2007, 131) writes that institutional *racism* occurs "when the administration or enforcement of the rules and procedures of a major social institution…is regularly distorted by the racial prejudice or bias of those who exercise authority within the institution. Institutional racism can exist even when the *content* of the rules and procedures of an institution, when viewed in the abstract, is perfectly just, provided there is pervasive racial bias in the *application* of those rules and procedures." Discrimination based on gender identity, sexual orientation, class, or religious affiliation, for example, could certainly be carried out in the same ways, rendering it "institutional" in Shelby's sense.

Perhaps we can find a way to utilize AI to combat institutional discrimination. We've already found ways to utilize AI to worsen it. Consider Amazon's recently-developed (and then abandoned) resume screener, which was designed to use AI to identify the top resumes in a large pool (Dastin 2018). Far from eliminating human prejudices, this tool automated them. It systematically favored men's over women's resumes, apparently because the algorithms it used were trained on data collected from Amazon's previous 10 years of applications—which were mostly from men (Dastin 2018). The automated screener favored resumes using certain words more often found in men's resumes, and actually downgraded resumes that contained the word "women's" (Dastin 2018). (Note, I often speak of gender in binary terms because that's how the research I discuss speaks of it. However, I don't mean to endorse a binary conception of gender. To the contrary—as Buolamwini and Gebru (2018, 6) also acknowledge—the reliance of (some)

existing research on a binary conception of gender represents another way in which dominant frames of thought can exclude.)

Assuming the training data was disproportionately male at least partly because of past unfairness—patriarchal norms discouraging women from working outside the home, popular sentiment that women weren't qualified to work in technical fields, unequal educational opportunities for women in computer science, etc.—the algorithm's reliance on this data is especially troubling. If used on real applications, Amazon's resume screener would ensure these historically-common forms of discrimination *continued* to disempower women: though women wouldn't be disallowed from applying to Amazon, their applications would be put at a significant disadvantage because of the biased system used to evaluate them.

We may also see such bias as contributing to what Iris Marion Young (2006; 2011) calls "structural injustice." According to Young (2006), different people occupy different positions within "social structures" (e.g., markets in certain goods and services), each position with its concomitant expectations, opportunities, advantages, and disadvantages. Social structures become sites of injustice when they systematically empower people in some positions while (and *by*) disempowering people others (Young 2006)—as, for example, men in a labor market using Amazon's resume screener would be systematically empowered because women were systematically disempowered. Moreover, Young (2006) argues that, by participating in a given social structure, we help perpetuate it; therefore, we are responsible for any injustice it creates.

Young's theory may also help us understand the gravity of Google's AI-powered photo sorter mis-identifying photos of black people as photos of "gorillas" (BBC 2015). The moral problem with this photo sorter is not (only) that its use would provoke offense, but that it would reinforce a mis-characterization of people of color (as less-than-human, or at least as less exemplary humans than white people) that's been invoked throughout history to justify horrific injustices like colonialism and slavery. Moreover, we have reason to believe that our present-day social structures bear the marks of these past injustices. For example, Anghie (2006) and Mutua (2000) argue that colonialism's central ideas and political objectives heavily influenced the development of international law and continue to structure global politics. If this is right, when we design and use technology that reflects and reinforces the central ideas and expectations underpinning these unjust structures, we perpetuate them and their constitutive injustice.

Epistemic Injustice

Miranda Fricker (2007, 1) defines epistemic injustice as an injustice "done to someone specifically in their capacity as a knower." More precisely, someone suffers "testimonial injustice" when others discount her credibility because of some prejudice (perhaps against her race or gender); and someone suffers "hermeneutical injustice" when "a gap in collective interpretive resources puts someone at an unfair disadvantage when it comes to making sense of their social experiences" (as, for example, women struggled to understand their own experiences of what we now call *sexual harassment* before this concept was developed) (Fricker 2007, 1). Further, hermeneutical injustice can result from what Fricker calls "hermeneutical marginalization"—when certain people are excluded from the collective processes by which a

society constructs the concepts necessary to interpret "some significant area(s) of social experience" (2007, 153).

The use of AI clearly implicates questions of epistemic (in)justice. Consider, again, Amazon's abandoned resume screener. It was trained on data representing members of a certain social group (men) to the exclusion of others (women). As a result, the AI system "learned" that what it meant to "have a good resume" was to "have a resume that looked like the resumes of previously successful men." In addition to perpetuating structural injustice, as outlined above, a society that adopted this system might begin to see this equation of being "good" or "qualified" with "being like previously successful men" as an objective truth. This could encourage testimonial injustice by creating (or reinforcing) the impression that women are not "good" or "qualified" by the standards of the tech industry, thereby undermining their credibility in that field—and perhaps in society more broadly, given the generally high esteem given to people deemed to have impressive technical skills. Moreover, if dominant ideas about what it means to be "good" or "qualified" are constructed on the basis of men's data, to the exclusion of women's, this is arguably an instance of hermeneutical marginalization.

Similarly, Buolamwini and Gebru (2018) evaluate three commercial programs that use machine-learning-based facial recognition software to classify images as "male" or "female." They find that these programs' accuracy varies substantially based on the gender and skin tone of the subjects being classified (Buolamwini and Gebru 2018). All three programs perform better on males than females and on lighter-skinned rather than darker-skinned people, and all perform worst on darker-skinned females (Buolamwini and Gebru 2018, 8).

This implicates questions of epistemic justice, because a society that relied heavily on the programs Buolamwini and Gebru (2018) discuss might come to identify looking "like a man" with looking like a *white* man. Certain ideas about what constituted a "masculine" trait (ideas that reflected the typical appearance of *white* men) could become generally accepted as "true," to the exclusion of alternative ideas (that might better reflect the typical appearance of men of color). Men of color might be seen (and made to see themselves) as deficient members of their gender; they may not match what their society has adopted as the paradigmatic picture of "man," because this picture was drawn to match the specifications of *white men*.

Again, this phenomenon could contribute to both kinds of epistemic injustice Fricker identifies: the testimony of men of color about their own gender identity or gender presentation may be discounted (testimonial injustice), and they may be denied a significant role in developing their society's collective understanding of what it means to look like or be a man (hermeneutical marginalization).

Moreover, given its relative inaccuracy when identifying women, people of color, and young people, law enforcement's use of facial recognition software could engender new kinds of discrimination against these demographic groups unknown to others (Buolamwini and Gebru 2018, 1-3). As Buolamwini and Gebru (2018, 1) speculate, "someone could be wrongfully accused of a crime based on erroneous but confident misidentification of the perpetrator from security video footage analysis." If women, people of color, and young people were systematically more vulnerable than others to such false accusations, this would arguably be an

example of structural injustice as Young understands it. Further, in a society in which these groups are already denied credibility or hermeneutically marginalized, the general public may be ill-suited to generate the conceptual resources necessary to understand this new form of discrimination, what makes it wrong, and what harms it causes, thereby furthering their hermeneutical marginalization.

Conclusions: AI as a Conduit for Human Power

The choice to delegate certain tasks to AI systems is sometimes described as a choice to take power *out of human hands*. But this, I propose, is a big mistake. If someone is denied a job because an AI system deems her resume inadequate, we'd be wrong to say she is subject "only" to the power of the AI system, or of its constituent algorithms, and not to the power of any human being. She is of course subject to the power of human beings—the people who wrote the code on which the AI system ran, the people on whose data the system's algorithms were trained, and the people who decided her resume would be evaluated by this particular AI system in the first place.

Some of this human power may have been exerted unwittingly. For example, those whose data was used to train the algorithm may not have known they were providing data for that purpose, or had a realistic opportunity to opt-out. (Though the fact that some people routinely harvest others' data without their genuine, informed consent itself results from human-created institutions and power structures.) And perhaps no one involved intended the specific outcome their collective action produced—to reject this particular candidate, or to discriminate against women candidates generally. Nonetheless, these human's actions, taken together, produced these outcomes. And, certainly, the decision to use AI to screen resumes was a human decision for which people should be held morally (if not politically or legally) accountable. Similarly, we'd be remiss to see the use of AI in law enforcement, state surveillance, or the targeting systems of autonomous weapons as anything but a particular way for some people to exert power over others. In these cases, human power may operate through a computer program, but it is a computer program written by humans, trained on human-created data, and put to work by some humans to monitor, regulate, control, and even exterminate others.

I don't mean to suggest that AI can never be a force for good. My point is that once we recognize it is a tool with which humans exercise power, rather than a *replacement* for human power, we must also recognize that its use raises questions of justice, as any other exercise of human power would. It is our responsibility as consumers, programmers, researchers, and people whose data are used to build AI systems to make sure these questions don't go unanswered.

References

Anderson, Elizabeth S. 1999. "What is the Point of Equality?" *Ethics* 109, 2: 287-337

Anghie, Antony. 2006. "The Evolution of International Law: Colonial and Postcolonial Realities." *Third World Quarterly* 27, 5: 739-53.

Aristotle. 2001. “Politics.” in Richard McKeon ed., *The Basic Works of Aristotle*. New York: Random House, Inc.

BBC. 2015. “Google apologises for Photo app’s racist blunder.” 1 July 2015. https://www.bbc.com/news/technology-33347866.

Beitz, Charles R. 1999. *Political Theory and International Relations: Revised Edition.* Princeton, N. J.: Princeton University Press.

Buolamwini, Joy and Timnit Gebru. 2018. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” *Proceedings of Machine Learning Research*: 81: 1-15.

Caney, Simon. 2005. *Justice Beyond Borders: A Global Political Theory*. Oxford: Oxford University Press.

Cohen, G.A. 1997. “Where the Action Is: On the Site of Distributive Justice,” *Philosophy & Public Affairs* 26, 1: 3-30

Dastin, Jeffrey. 2018. “Amazon scraps secret AI recruiting tool that showed bias against women.” *Reuters.* October 9, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Miller, David. 2007. *National Responsibility and Global Justice.* Oxford: Oxford University Press.

Mutua, Makau. 2000. “What is TWAIL?” with comment by Antony Anghie. *Proceedings of the Annual Meeting (American Society of International Law)* 94: 31-40.

Nussbaum, Martha. 2000. *Women and Human Development: The Capabilities Approach.* Cambridge: Cambridge University Press.

Nussbaum, Martha. 2003. “Capabilities as Fundamental Entitlements: Sen and Social Justice.” *Feminist Economics* 9, 2/3: 33-59.

O’Neill, Onora. 2005. “The Dark Side of Human Rights.” *International Affairs (Royal Institute of International Affairs 1944-).* 81, 2: 427-39.

Okin, Susan Moller. 1989. *Justice, Gender, and the Family.* New York: Basic Books, Inc.

Rawls, John. 1999. *A Theory of Justice: Revised Edition.* Cambridge, Mass. The Belknap Press of Harvard University Press.

Sen, Amartya. 1980. "Equality of What?" in Sterling M. McMurrin ed., *The Tanner Lectures on Human Values, Volume 1*. Cambridge: Cambridge University Press.

Shelby, Tommie. 2007. "Justice, Deviance, and the Dark Ghetto." *Philosophy & Public Affairs.* 35, 2: 126-60.

Shue, Henry. 1996. *Basic Rights: Subsistence, Affluence, and U.S. Foreign Policy: Second Edition*. Princeton, N.J. Princeton University Press.

Young, Iris Marion. 2011. *Responsibility for Justice.* Oxford: Oxford University Press.

Young, Iris Marion. 2006. "Responsibility and Global Justice: A Social Connection Model." *Social Philosophy and Policy* 23, 1: 102-30.

Ypi, Lea. 2013. "Cosmopolitanism Without If and Without But." in Gillian Brock ed., *Cosmopolitanism versus Non-Cosmopolitanism: Critiques, Defenses, Reconceptualizations*. Oxford: Oxford University Press.

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

Suggested Reading

<u>Historical Works on Justice</u>

Aristotle. 2001. "Politics." in Richard McKeon ed., *The Basic Works of Aristotle*. New York: Random House, Inc.

Douglass, Frederick. 2014. *My Bondage and My Freedom.* New Haven, Conn.: Yale University Press.

Du Bois, W.E.B. 1999. *The Souls of Black Folk*. Norton Critical Edition, Henry Louis Gates Jr. and Terri Hume Oliver, eds. New York, W.W. Norton & Company.

Du Bois, W.E.B. 1998. *Black Reconstruction in America, 1860-1880.* New York: The Free Press.

Hobbes, Thomas. 1994. *Leviathan*, Edwin Curley ed., Indianapolis, Ind.: Hackett Publishing Company.

Locke, John. 1988. “The Second Treatise of Government: An Essay Concerning the True Original, Extent, and End of Civil Government.” in Peter Laslett ed., *Locke: Two Treatises of Government.* Cambridge: Cambridge University Press.

Plato. 1997. “Republic.” in John M. Cooper ed., *Plato: Complete Works.* Indianapolis, Ind.: Hackett Publishing Company.

Rousseau, Jean-Jacques. 1997. “Discourse on the Origin and Foundation of Inequality Among Men.” in Victor Gourevitch ed., *Rousseau: The Discourses and Other Early Political Writings.* Cambridge: Cambridge University Press.

Rousseau, Jean-Jacques. 1997. “Of the Social Contract.” in Victor Gourevitch ed., *Rousseau: The Social Contract and Other Later Political Writings.* Cambridge: Cambridge University Press.

Wollstonecraft, Mary. 2010. *A Vindication of the Rights of Woman with Strictures on Political and Moral Subjects.* Cambridge: Cambridge University Press.

Theories of Justice (Contemporary, General)

Nozick, Robert. 1974. *Anarchy, State, and Utopia.* New York: Basic Books.

Nussbaum, Martha. 2000. *Women and Human Development: The Capabilities Approach.* Cambridge: Cambridge University Press.

Okin, Susan Moller. 1989. *Justice, Gender, and the Family.* New York: Basic Books, Inc.

Pettit, Philip. 2012. *On the People’s Terms: A Republican Theory and Model of Democracy.* Cambridge: Cambridge University Press.

Rawls, John. 1999. *A Theory of Justice: Revised Edition.* Cambridge, Mass. The Belknap Press of Harvard University Press.

Walzer, Michael. 1983. *Spheres of Justice: A Defense of Pluralism and Equality.* New York: Basic Books, Inc.

Structural Injustice (Contemporary)

Julius, A.J. 2006. “Nagel’s Atlas.” *Philosophy & Public Affairs.* 34, 2: 176-92. (NB: This is a reply to Thomas Nagel. 2005. “The Problem of Global Justice.” *Philosophy & Public Affairs* 33, 2: 113-47.)

Lu, Catherine. 2011. "Colonialism as Structural Injustice: Historical Responsibility and Contemporary Redress." *Journal of Political Philosophy* 19, 3: 261-81.

Young, Iris Marion. 2011. *Responsibility for Justice.* Oxford: Oxford University Press.

Young, Iris Marion. 2006. "Responsibility and Global Justice: A Social Connection Model." *Social Philosophy and Policy* 23, 1: 102-30.

Theorizing Power and Exclusion (Contemporary)

Fanon, Frantz. 2008. *Black Skin, White Masks*. Richard Philcox, trans. New York: Grove Press.

Fanon, Frantz. 2004. *The Wretched of the Earth*. Richard Philcox, trans. New York: Grove Press.

Foucault, Michel. 2008. *The Birth of Biopolitics: Lectures at the Collège de France, 1978-1979.* Graham Burchell, trans. New York: Palgrave Macmillan.

Foucault, Michel. 1995. *Discipline and Punish*. Alan Sheridan, trans. New York: Vintage Books.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

MacKinnon, Catharine A. 2005. *Women's Lives, Men's Laws.* Cambridge, Mass.: The Belknap Press of Harvard University Press.

MacKinnon, Catharine A. 1989. *Toward a Feminist Theory of the State.* Cambridge, Mass.: Harvard University Press.

Malcolm X. 1965. "Appeal to African Heads of State." in George Breitman ed., *Malcolm X Speaks: Selected Speeches and Statements.* New York: Grove Press.

Malcolm X. 1965. "The Black Revolution." in George Breitman ed., *Malcolm X Speaks: Selected Speeches and Statements.* New York: Grove Press.

Mills, Charles. 1997. *The Racial Contract.* Ithaca, N.Y.: Cornell University Press.

Lukes, Steven. 1974. *Power: A Radical View.* London: The Macmillan Press, Ltd.

Pateman, Carole. 1988. *The Sexual Contract.* Stanford, Calif.: Stanford University Press.

Shelby, Tommie. 2007. "Justice, Deviance, and the Dark Ghetto." *Philosophy & Public Affairs.* 35, 2: 126-60.

Shelby, Tommie. 2005. *We Who Are Dark: The Philosophical Foundations of Black Solidarity*. Cambridge, Mass.: The Belknap Press of Harvard University Press.