Political Philosophy in the AI Ethics Classroom

Abstract

This paper defends two main claims. First, that political philosophy deserves a central place in AI Ethics' curricula. This is a claim about the content of the AI Ethics class. The second claim is about the form of the AI Ethics class: namely, that considerations originating in political philosophy must inform the way in which AI Ethics is taught. The basic idea animating both claims, is that AI has powerful political implications and that preparing students to navigate these implications, requires paying close attention to both the cognitive and practical learning goals of the AI Ethics course.

Introduction

AI Ethics has been a standard course offering of many college philosophy departments for several years now. It is also increasingly becoming part of the college computer science curricula, sometimes in the form of required or elective courses, and more often in the form of ethics units embedded in other, technical courses. It is in these courses and units that students are taught to ask—and to attempt to answer—normative questions about artificial intelligence.

The emergence of the field—as a subdiscipline one can study at university and as an area of academic and industry research—can be viewed as a sign of progress.[i] Indeed, since late 2022, when ChatGPT was released and became the "fastest-growing consumer application in history" (Gordon, 2023), the push for AI Ethics courses and units at the college level has only increased, as the importance of preparing citizens, and future technologists, to responsibly navigate an AI future has felt more pressing than ever.

However, ethics is not the only discipline that is concerned with normative questions. Political philosophy is another such discipline. Yet, courses on the political philosophy of AI are hardly a standard offering, even among those colleges leading the pack when it comes to connecting AI and the humanities. While any account of the distinction between the two subdisciplines is bound to court controversy, it is reasonable to say that where ethics encourages normative questions about personal and interpersonal matters, political philosophy asks questions about our collective lives and *political* systems.

I argue that several of the most important normative questions that are discussed in standard AI Ethics classes are questions that educators cannot hope to help students answer well without introducing some political philosophy. In saying this, I am advocating for a central role of political philosophy in the AI Ethics classroom. For ethicists teaching AI Ethics classes, I do not expect this to sound particularly new or controversial. However, AI Ethics courses are not always taught by ethicists — they are quite often taught by instructors in other disciplines, like Computer Science. My purpose here, then, is to convince these instructors that political philosophy deserves a central place in their AI Ethics courses, because AI has powerful political implications. I make my case for this claim, which is a claim about the *content* of AI Ethics courses, in the first section of this article. As I argue in the second section, the truth of this claim has implications for the way AI Ethics class should be taught. The second section, then, will advocate for a claim about the proper *form* of AI Ethics classes.

1. What We Should Teach, When We Teach AI Ethics?

I make my case for the first claim by taking up two central topics in AI Ethics: AI bias and privacy. I explain why each topic raises important, distinctly *political* questions and argue that, consequently, to prepare students to answer them well, instructors need to introduce some political philosophy into the AI Ethics classroom.

1.1. AI bias

To speak of AI "bias" in an ethics class is to name a feature of an AI tool that is morally bad.[ii] AI bias is a topic that deserves, and has been rightly accorded, a central place in the AI Ethics curriculum. When tackled in the college-level classroom, the topic invites students to grapple with the fact that AI models are only as good as the data they are trained on and that, far from approximating a "God's eye view" of their subject matter, data can reflect a partial, contestable, or even outright inaccurate, view of things. Investigating the sources of algorithmic bias thus demonstrates the dangers of a naive trust in numbers. When I teach AI Ethics, I am especially interested in helping my students not only understand the common-refrain in computer science of "garbage-in, garbage-out," but to ensure they are able to recognize the different ways in which datasets might be appropriately judged to be "garbage." I want them to distinguish, for instance, between bias that can be traced to the fact that the data used to train a model is not sufficiently diverse (as when datasets used to develop facial recognition models include mainly white faces[iii]), from bias that originates in the prejudice-induced errors of the people collecting the training data (as when that data consists of past decisions about who to admit to medical school, but those decisions reflect the admissions committee's racial prejudice[iv]).

The ability to identify different sources of potential bias enables students to read empirical research critically, and will serve many of them well in their future careers. Although not all students will go on to work as technologists, a great many will be tasked with handling (often very large) data sets. However, to focus only on the causes of bias in AI is to take for granted that we all know when, and why, AI tools are biased in the relevant sense. If we take these things for granted, we risk leaving students unprepared to accurately identify new and more subtle forms of bias, and to evaluate the various methods we might embrace to mitigate them.

Consider Gabriel Johnson's (2020, 9947-9948) discussion of a hypothetical algorithm designed to classify whether individuals are good or bad at computers. The model ends up predicting that elderly people are significantly less likely to be good at computers. Here, the model's predictions mirror a stereotypical belief about the capacities of elderly people and, for this reason, Johnson calls the model "problematic" (9948). However, as the hypothetical tool is described, the model's training data need not be inaccurate or unrepresentative. Rather, the data may accurately represent facts about the distribution of computer-skills in a society that is shaped by injustice: in an ageist society, elderly people might be statistically more likely to be bad at computers because they have been unfairly denied opportunities to gain computer skills. As Johnson notes, the example demonstrates an important, third way in which AI bias can arise; in a world shaped by injustice, datasets do not need to be unrepresentative or inaccurate to produce outputs that exemplify prejudicial stereotypes. The problem with Johnson's discussion, though, is the suggestion that whether an AI model is biased in a way that is morally problematic, is one we can answer without paying attention to the way the algorithm will be used. After all, Johnson calls the model problematic, without describing how, exactly, it is being used in the hypothetical context.

However, to determine whether the model really is biased, we need to consider how it is being used.[v] Suppose the model was being used to distribute invitations to a free computing course. If this is the case, then the fact that the model demonstrates unequal false-positive rates between elderly people and younger people is *not* a cause of concern. Indeed, that the rate of false positives is higher for elderly people can be understood as corrective to the historically unfair distribution of opportunities for technical training. If, on the other hand, the model is being used to screen job applicants, such that older candidates are less likely to get invited to interview than younger candidates, then we would have reason to be concerned. That's because, if age correlates negatively with computer skills (and computer skills are relevant to determining qualifications), then elderly people will be denied opportunities based on skills that have no had equal opportunities to develop. This seems unfair and will be especially objectionable if computer skills can be gained easily on the job.

This suggests that the historical context in which AI is being used, together with facts about whether it is being used to distribute benefits or burdens, matters a great deal to whether an AI tool should be evaluated as problematically biased. Recall the famous COMPAS investigation (Angwin et al. 2016) — one of the first cases that forced the issue of algorithmic bias into public consciousness. COMPAS is used to help judges determine who is eligible for parole and bail by predicting who is at higher risk of re-offending. The investigation found that Black and Hispanic men experienced a much higher rate of false "high-risk" classifications than white men. That was cause for alarm because it meant that the burden of a false positive was being disproportionately shouldered by the group that had *already* been disadvantaged by a host of historical injustices in the U.S. Here, the pattern of error does not rectify, but reinscribes historical injustices.

What these examples indicate is that students need to grapple with questions about equality of opportunity, and hence questions of justice, to responsibly evaluate claims of AI bias. This, in turn, requires introducing some political philosophy into the AI Ethics classroom. Students need to be able to distinguish between different kinds of equality of opportunity (to distinguish between formal equality of opportunity, for instance, and substantive equality of opportunity), their rationales, and the circumstances in which one might endorse one over the other (see also Castro, O'Brien, and Schwan, 2023). In my discussion of the hypothetical tool described by Johnson, formal equality of opportunity might have been satisfied (the training data used to create the model didn't need to include an 'age' column[vi]), but a commitment to substantive equality was what drove my moral evaluations: when the tool was being used to distribute computer training opportunities, it was being used to help realize substantive equality of opportunity. When it was being used to distribute job opportunities, it was violating substantive equality of opportunity.

Focusing students' attention on these different ways of operationalizing a commitment to egalitarianism is also necessary if students are to appropriately evaluate the variety of technical responses available to the problem of biased AI. These technical responses, also known as *statistical fairness criteria*, are constraints that can be built into an AI model in the training phase. For instance, when training a model that will be used to predict the likelihood of a person being arrested, we might ask that the model optimize for accuracy while also ensuring equal rates of false positives and negatives (i.e., "equalized odds") for different racial groups. A variety of statistical fairness criteria have been proposed in the machine learning community in the last ten years or so. The problem is that no single criterion is a sufficient, or even necessary, means for ensuring fairness across all contexts. It is not even the case, moreover, that different measures help realize *different ways* of operationalizing a commitment to equality. As some philosophers have recently been at pains to establish, it is not the case that there is one measure (e.g., so-called "fairness through awareness") that is relevant to formal equality of opportunity, another relevant to substantive equality (e.g., equalized odds), and yet another that is relevant to the particular flavor of substantive equality that is luck egalitarianism (e.g., so-called "counterfactual fairness") (Castro, O'Brien, and Schwan, 2023). For citizens to know whether a particular statistical fairness measure is evidence for the fairness of a given AI tool—and for technologists to know which measure is relevant to what they are designing —requires careful attention to the prospective use to which the tool will be put, and how the broader social context might give reason for opting for one kind of equality over another.

Among the concerns relevant to all of this, moreover, will be the values at stake when deciding to operationalize for anything other than predictive accuracy.[vii] Recall Johnson's computer-skills prediction tool which, I suggested, would violate a commitment to substantive equality if used screen job applicants. Even in this context, there may be room for disagreement about how to weigh the value of substantive equality of opportunity against the value of ensuring that as many qualified candidates get the job as possible. If on-the-job training would take considerable time and come at the cost of some other important value that the firm is in service to — say, fighting climate change, curing a neglected disease, or economic interventions aimed at benefitting the poorest groups in society — it might be all-things-considered acceptable for the tool to violate substantive equality of opportunity. To prepare students to responsibly engage with the question of whether the use of this kind of algorithmic tool is justified, then, students need to practice weighing the value of equality, against the values of (for instance), improving the situation of the poorest or averting a catastrophe for all. Reasoning about such trade-offs is a central task of political philosophy.

1.2. Data privacy

AI models are almost always trained on data that are supplied by the behavior, or intellectual output, of people other than those who design and profit from them. Consider, for instance, the way Large Language Models (LLMs) like Google's Gemini or OpenAI's GPT models are trained on text produced by other people on sites like (to name just a few) Wikipedia, Reddit, news websites, and social media. Or, to take an example that predates LLMs, consider the machine learning algorithms, trained on information about people's online behavior, to service the business of targeted advertising — the business that underwrites most of the modern internet. AI's need for data, and the ways in which it is currently sourced by developers, raises a host of normative questions. Are AI companies violating our *privacy* when they collect data about our online behavior, or when they use that data to make detailed inferences about us? Are AI companies *stealing* when they use data that is produced by other people? These are some of the major questions students in my AI Ethics classes are raising. As I argue, the nature of big data and the broader data economy that underwrites the current AI boom mean that we cannot prepare students to successfully navigate these issues without introducing some political philosophy.

Many people think that when companies collect data about our online behavior, they violate our privacy. However, to say that something violates privacy is often a way of just saying that it is morally bad, rather than explaining why it is morally bad. When pressed to say what privacy is or why it is valuable, for instance, few students know how to proceed. The task for students in an AI Ethics class, then, is to try to identify which features they are responding to when they assert that certain data practices violate privacy, and to explain why those features give us normative reasons. One tempting way to do this is to focus on the issue of consent and, therefore, the value of personal autonomy. Respecting privacy, on this approach, would seem to be about respecting autonomy: if I do not consent to being surveilled online but am surveilled anyway, then my autonomy has been disrespected. This is the approach that many students embrace, and it is the one they are likely to see endorsed by congress people and other advocacy groups. The Kids Online Safety Act, for instance, does not prevent companies from collecting data about minors, but it does have provisions designed to strengthen parental consent.

In the age of big data, however, improving consent mechanisms does not ameliorate privacy concerns. That's because, even if I do not consent to share my data, companies are still able to learn a great deal about me in a way that seems to violate my privacy, provided that a sufficient number of other people do consent to sharing their data. That this is possible is most easily understood in the context of online social networks (see Barocas & Nissenbaum 2014; see also Viljoen 2021). Although one person might refuse to share their data with a social media company, undisclosed information about that person can still be inferred from information that their online "friends" have chosen to disclose. Studies have shown, for instance, that detailed information about a person's education, as well as their sexuality, can be inferred from features of the social networks they maintain online (Mislove et al. 2010; Jernigan and Mistree 2009). Significantly, though, these kinds of inferences do not always require information about explicit social connections, and so are possible outside of social networking contexts. In the widely discussed example, Target learnt to predict whether a customer was pregnant, with some reliability, based on information about their shopping habits (Duhigg 2012). They were able to do this, moreover, only because a very small proportion of Target's customers has decided to disclose that they had recently given birth. All this means that an emphasis on consent can leave students with an inaccurate understanding of big data's challenge to privacy and how it might be successfully addressed (cf. Zuboff 2019).

Here, social theories of privacy can help. Helen Nissenbaum's (2009) theory of contextual integrity is a prominent theory of this kind, and is a staple of many digital ethics curricula. On this view, we speak of privacy violations to refer to violations of contextually determined social norms. So, for instance, there is a norm according to which patients can share deeply personal information with their doctors (something that would be inappropriate in other contexts), but which prohibits doctors from sharing that information with anyone else. Contextual norms like this one earn their keep in virtue of the fact that when they are upheld by a society, they enable us to achieve *other* values. So, the requirement of doctor-patient confidentially means that everyone can speak freely with their doctor, and which in turn improves public health outcomes.

Applied to the question of digital surveillance, Nissenbaum's views force us to ask: "what social value(s) are undermined (or risked) by the sort of ubiquitous data collection that characterizes the contemporary digital sphere, and which makes AI (in all its guises) possible?" When we try to answer this question, we land ourselves in the domain of political philosophy. While my data is not particularly valuable on its own, it is valuable when it is part of a much larger data set. Those who control that data set will be able to draw a huge range of detailed inferences about an indeterminately large range of people. This is an enormous asymmetry in knowledge and, consequently, in power — an asymmetry that students need to be encouraged, and enabled, to interrogate. After all, while it might not seem like much is at stake when, for instance, AI enables Google to target someone with ads for bubble tea when they are craving sugar, this feature of big data—a feature that enables private AI companies to know a great deal about us, and do what they wish with that knowledge—actually makes everyone vulnerable to potentially *dominating* power (Pettit, 1997; Roberts, 2015). Consider, for instance, the fact that, in 2023, a Colorado Catholic group purchased cell phone and app data that enabled it to locate and identify gay priests in their diocese (Boorstein and Kelly 2023). The data practices that make AI possible are precisely the same practices that enabled those gay priests to be identified, companies to predict who is susceptible to abusive payday loans, or who might be considering an abortion. In each case, the knowledge big data and AI confers enables some people to wield tremendous power over others, in way that seems to offend a commitment to equality.

If students are to be empowered to be ethical change-makers when it comes to the future of AI, they need to be capable of going beyond general claims about the violation of privacy violation—claims that focus on the transactions between individuals and AI companies—and to learn to evaluate the power asymmetries that collectively, our data has enabled. Consequently, students need to be empowered think about our rights, as a collective, to shape the uses to which our data is put. All this casts a new light on the idea that concerns about data collection can be properly addressed by ensuring that AI companies financially compensate users when they collect and/or use their data. For one thing, if we conceive of data as a kind of property, then no person's data is going to be worth very much, economically speaking (cf. Benn and Lazar, 2022). More importantly, however, it is not even obvious that monetary compensation for data is the appropriate response to the worry. After all, while compensation would seek to address one aspect of the power asymmetry (namely, the economic aspect,), it would leave untouched the distinctly epistemic aspect of that asymmetry.

2. How We Should Teach, When We Teach AI Ethics?

In this section, I argue that the political implications of AI should impact not just *what* is taught in the AI Ethics classroom, but also the way such classes are taught. That's because, while political philosophy can give students the cognitive resources they need to think critically about the political consequences of AI, more than cognitive resources are needed if students are to be empowered to impact the way these consequences actually unfold. For this, a set of civic skills is necessary – skills that go beyond the ability to produce and disseminate knowledge, and which enable citizens to engage with each other in particular ways. If AI Ethics courses are going to enable students to foster these skills, then instructors ought to pay attention to how they teach, in addition to what they teach.

Although it would go beyond the scope of this paper to defend a fully fleshed-out account of civic virtue, here I focus on two component skills that, I hope, will strike most readers as uncontroversial. First, civic virtue requires that people know how to express their reasons effectively to a variety of other people, and to interpret the views of those other people charitably. Second, they must be comfortable with disagreement; to able to see it as something that can be productive and good for the collective, and as something that should not undermine our respect for those with whom we disagree. Together, these are skills that students need if they are to be disposed to raise, and then reason about, the social place of AI with a wide variety of other people.

How can AI Ethics classes be designed to enable students to hone these skills? The first step is to communicate, to students, that the development of these skills is one of the course's key learning goals. To increase student buy-in, this can be done by connecting the importance of these skills to the broader, content-based learning objectives of the course – in other words, by making explicit that the class is not just oriented towards understanding that AI raises difficult political questions, but towards learning how to answer those questions with other people. The hope, in foregrounding these practical goals, is encourage students to articulate their agreements and disagreements with each other during lessons. The next step, of course, is to design lessons so that students are given ample opportunity to reason with each other, about how the political consequences of AI should unfold. In what comes, I sketch three ways of doing this, all of which I've experimented with in my own teaching.

My first suggestion concerns a preliminary discussion prompt that encourages students to begin seeing digital technology, in general, as political. "Technological solutionism" (Morozov 2013) is a concept I have introduced to most of my classes on digital ethics. It refers to a tendency to uncritically embrace technological solutions to highly complex, often collective, problems. When discussing technology that provide real benefits to some, the concept provides a useful frame for student disagreement about how inequality ought to be tackled. Consider, for instance, the Moxie robot that is designed to help kids that struggle with emotional skills. The robot is significantly cheaper than in-person therapy, and so meets the needs of children with special needs, whose families cannot afford in-person therapy. However, it may do so less effectively than in-person therapy. Is Moxie an example of technological solutionism? On the one hand, society could improve access to in-person therapy for children. A world in which this happened would be a world in which all kids got what they deserved: the attention and care of a real person with expertise that can help them. But creating that world is difficult and will inevitably take a long time. On the other hand, Moxie is providing real benefits to children and families right now. Should meeting their needs be forced to wait on a collective fight for fair access to therapists (a fight which will certainly be drawn out, and is not guaranteed to succeed)? When students are asked to tackle these questions together, reasonable disagreement is hard to avoid. Many students are likely to have family members or friends with the sort of special needs Moxie is designed to meet, and some of these may not have been able to easily access in-person therapy. An issue like this can thus enable students to mine their own experience for considerations that bear on our evaluation of the AI, providing an opportunity for students to learn from each other and grapple with the question of how to acknowledge and accommodate a variety of conflicting, albeit reasonable, perspectives. More fundamentally, it's an issue that illuminates the political consequences of even seemingly uncontroversial technological interventions.

My second suggestion concerns AI bias, and invites students to connect the political questions it raises to issues they are likely to already have some stake in. Instructors ask students to imagine that their university is developing an algorithmic tool designed to improve admissions decisions. One version of the tool uses student GPA and graduation data from across the country, to predict who is most likely to succeed at university. A second version does the same, but then artificially boosts the results for students from lower socio-economic backgrounds, to ensure that a greater proportion of such students are admitted. When students are then tasked with comparing the two tools, they should, themselves, arrive at two, competing conceptions of equality: formal and substantive equality of opportunity.[1][viii] While the first tool realizes the former conception, the second tool arguably works towards substantive equality of opportunity. Deciding which tool is appropriate thus requires that student's reason with each other about the proper goals of higher education in our society. If one thinks that higher education is primarily a vehicle of opportunity, then one will be inclined to favor the second version of the tool. On the other hand, if one is inclined to think that primary purpose of higher education is to produce citizens that will both produce new scientific knowledge and contribute to economic growth, then perhaps the first tool will look preferable. These options

[1]

are, of course, not exhaustive. For instance, students should be encouraged to also consider the ways in which different varieties of diversity are liable to confer considerable epistemic benefits, too. However, the general idea is to illuminate the fact that evaluation the predictive tool cannot be undertaken without considering the proper aims of higher education. Such considerations take us into the domain of politics, however, and so the task of designing an algorithmic tool to help with admissions cannot be straightforwardly outsourced to computer scientists. Further disagreement can then be elicited by asking students to consider how their preferences would change if the tool were designed to determine admissions to a high-stakes research program for the public good – like one that aims to develop green-energy solutions that are necessary to prevent global suffering due to climate change, for instance. If the stakes are sufficiently high, might we have greater reason to optimize for predictive accuracy?

My third and final suggestion, provides a way for students to practice the relational skills necessary for civic virtue, via a kind of role play where students experiment with models of data self-governance. Students are put into groups and instructed to imagine that they are members of a new, self-governing social network. As such, they must decide, collectively, how the network's data will be used. Will they, as a collective, decide to sell their data, to help maintain the network, to make it more accessible, or to otherwise improve its functionality? If so, what kind of data will they make available and who are they willing to sell it to? To scaffold the exercise, instructors can first give groups the opportunity to decide what kind of network they will form – a dating app, for instance, a local neighborhood network, or an Instagram-like platform. This decision will help determine the content of the data the network is likely to generate, and so give students some sense of the inferences that would be enabled by those with access to that data. Then, instructors introduce certain data-sharing proposals from different bodies – commercial, civil society, and governmental. Once presented with several proposals, groups need to consider the possible consequences of the agreements they make, reach a consensus about what to do, and then justify their decisions to the rest of the classroom. The goal is to have students practice making decisions about how to balance the benefits and risks of data collection in ways that are publicly justifiable. Doing this well, moreover, is more likely when a diverse group of individuals is included in that decision-making process. After all, while sharing data about one's health or romantic life might appear to court no serious risk from the perspective of some individuals, from the perspective of others it may appear obviously too risky to truly countenance.

Conclusion: AI *in* education

In closing, I draw out one implication of my argument so far, for how we should be thinking about the educative uses of AI at the college level. Since the release of ChatGPT, there has been a lot of anxiety amongst those teaching at the tertiary level about how to how to prevent students from relying excessively on Large Language Models (LLMs). This is only partly a worry about cheating – it is also a worry that students, who were quick to appreciate the capabilities of tools like ChatGPT, will grow skeptical of the value of the learning goals that their college instructors set for them. Why learn how to write a philosophy argument or analyze a work of literature, for instance, if AI could (or soon would) do that for them? What would have once counted as "cheating" was coming to look, at least to some students, like deciding to use an abacus over a calculator.

In response, many instructors have had to rethink how they communicate the value of their discipline-specific learning goals to their students. In doing so, the hope is to make clear, to students, that to rely on AI when completing their coursework would mean cheating themselves of one of the intrinsically valuable opportunities that college provides. The challenge of communicating this fact is hardly a new one. However, things have become significantly more complicated with the rise in popularity of LLMs. That's because, depending on how instructors understand and articulate the learning goals of their courses, there might be some reason to doubt that these goals *cannot* be achieved by interacting with AI alone.[ix] And, if that's the case, why go to university at all? For instance, suppose one agrees with Stanley Fish (2008) that college instructors should only empower students to produce and disseminate disciplinary knowledge. If this is right, then the value of a college education is going to be seriously at risk. LLMs can, for instance, explain philosophical concepts clearly and accurately. They can also be prompted to explain objections to arguments, and so can be a helpful resource for understanding the diversity of views that have been defended with respect to a given philosophical topic. There is reason to think, moreover, that LLMs can help people not only understand and so disseminate existing knowledge, but ultimately produce *new* knowledge, too.

Of course, nearly everyone is now familiar with the ways in which ChatGPT makes wildly stupid mistakes. Accordingly, it might seem overblown to claim that LLMs can help students foster the cognitive skills of a discipline like philosophy, or to even push that discipline forward. However, it is possible that these technical limits will be overcome in time. Moreover, if we focus on the fact that these systems cannot, in virtue of their present limitations, enable our students to achieve the learning goals we set for them, then we sell ourselves significantly short. Indeed, we risk implying that when these systems are up to snuff, a college-level education may not require a college classroom at all. If the arguments of this paper are correct, however, then this is an implication that instructors of AI Ethics courses have strong reason to reject. A central of goal of these classes is the production, in students, of the knowledge *and* practical skills that will enable them to work with others to help shape the political consequences of AI. These are skills that simply cannot be fostered by engaging with a sophisticated LLM. That's because, as I have argued, AI raises questions about the proper goals of our institutions, about the just distribution of benefits and burdens, and about the legitimacy of power relationships. These are questions that need to be answered collectively, and so call for the development of the deliberative, relational skills that will enable students to reason effectively with their fellow citizens. When instructors make this goal explicit, they foreground a function of higher education that cannot be achieved by interacting with even a sophisticated AI.

References

Angwin, Julia, Jeff Larson, Mattu Surya, and Lauren Kirchner. 2016. "Machine Bias." ProPublica. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, Solon, and Helen Nissenbaum. 2014. "Big Data's End Run around Anonymity and Consent." In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 44–75. Cambridge University Press.

Benn, Claire, and Seth Lazar. 2022. "What's Wrong with Automated Influence." *Canadian Journal of Philosophy* 52 (1): 125–48.

Boorstein, Michelle, and Heather Kelly. 2023. "Catholic Group Spent Millions on App Data That Tracked Gay Priests." *Washington Post*, May 2, 2023. https://www.washingtonpost.com/dc-md-va/2023/03/09/catholics-gay-priests-grindr-data-bishops/.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Conference on Fairness, Accountability and Transparency*, 77–91. PMLR. http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline&ref=akusion-ci-shi-dai-bizinesumedeia.

Castro, Clinton, David O'Brien, and Ben Schwan. 2023. "Egalitarian Machine Learning." *Res Publica* 29 (2): 237–64. https://doi.org/10.1007/s11158-022-09561-4.

Duhigg, Charles. 2012. "How Companies Learn Your Secrets." *The New York Times Magazine*, February 16, 2012.

Fleisher, Will. 2023. "Algorithmic Fairness Criteria as Evidence." SSRN Scholarly Paper. Rochester, NY. https://papers.ssrn.com/abstract=3974963.

Gordon, Cindy. n.d. "ChatGPT Is The Fastest Growing App In The History Of Web Applications." Forbes. Accessed April 30, 2024. https://www.forbes.com/sites/cindygordon/2023/02/02/chatgpt-is-the-fastest-growing-ap-in-the-history-of-web-applications/.

Jernigan, Carter, and Behram F. T. Mistree. 2009. "Gaydar: Facebook Friendships Expose Sexual Orientation." *First Monday* 14 (10).

Johnson, Gabbrielle M. 2020. "Algorithmic Bias: On the Implicit Biases of Social Technology." *Synthese* 198 (10): 9941–61.

Lowry, Stella, and Gordon Macpherson. 1988. "Blot on the Profession." *British Medical Journal* 296 (657).

Mislove, Alan, Bimal Viswanath, Krishna P. Gummadi, and Peter Druschel. 2010. "You Are Who You Know: Inferring User Profiles in Online Social Networks." In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 251–60. WSDM '10. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1718487.1718519.

Morozov, Evgeny. 2013. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: PublicAffairs.

Nissenbaum, Helen. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford: Stanford University Press.

Ochigame, Rodrigo. 2019. "How Big Tech Manipulates Academia to Avoid Regulation." The Intercept. December 20, 2019. https://theintercept.com/2019/12/20/mit-ethical-ai-artificial-intelligence/.

Pettit, Philip. 1997. *Republicanism: A Theory of Freedom and Government*. Oxford: Oxford University Press, Incorporated. http://ebookcentral.proquest.com/lib/georgetown/detail.action?docID=3052800.

Roberts, Andrew. 2015. "A Republican Account of the Value of Privacy." *European Journal of Political Theory* 14 (3): 320–44.

Viljoen, Salome. 2021. "A Relational Theory of Data Governance." *Yale LJ* 131:573.

Wang, Angelina, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2024. "Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms That Optimize Predictive Accuracy." *ACM J. Responsib. Comput.* 1 (1): 9:1-9:45.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism*. First edition. New York: Public Affairs.

[i] For a more critical different perspective see Ochigame (2019).

[ii] In statistics and machine learning, the traditional sense of "bias"—namely, statistical bias—means something very different from the sense at issue here. (See Barocas and Nissenbaum 2014).

[iii] In statistics and machine learning, the traditional sense of "bias"—namely, statistical bias—means something very different from the sense at issue here. (See Barocas and Nissenbaum 2014).

[iv] For example, see Lowry and Macpherson (1988).

[v] For more on these points, and on the significance of statistical fairness measures in light of them, see Fleisher (2023).

[vi] The standard way to put this is to say that age has been *redundantly encoded*.

[vii] In the machine learning literature, this is often described in terms of a trade-off between fairness and accuracy. Of course, accuracy is not itself a social or political value, at least not directly so. Students should, instead, be taught to ask about the value to which accuracy is in service.

[viii] For an appropriate accompanying text, see Castro, O'Brien, and Schwan (2023).

[ix] Whether college would still be worth the large-scale investment that it currently invites would depend on the value of everything that happens at college *outside of* the classroom.