APRIL 2022

JUSTICE, HEALTH, AND DEMOCRACY IMPACT INITIATIVE &
CARR CENTER FOR HUMAN RIGHTS POLICY

A More Equal Future?

Political Equality, Discrimination, and Machine Learning

Technology & Democracy
Discussion Paper

t

A More Equal Future?
Political Equality, Discrimination, and Machine Learning

Joshua Simons
Postdoctoral Fellow in Technology and Democracy
Edmond J. Safra Center for Ethics
Carr Center for Human Rights Policy
Harvard University

Eli Frankel
Undergraduate Fellow
Edmond J. Safra Center for Ethics
Harvard University

Series edited by **Joshua Simons**

The content of this piece is based on work found in the Author’s upcoming book, *Algorithms for the People: Democracy in the Age of AI*, to be published through Princeton University Press in fall 2022.

Harvard University
April 2022

JUSTICE, HEALTH
& DEMOCRACY
IMPACT INITIATIVE

HARVARD Kennedy School
CARR CENTER
for Human Rights Policy

The Justice, Health, and Democracy Imapct Initiative is a multi-institutional collaboration between New America, the Brown University School of Public Health, and the Edmond J. Safra Center for Ethics at Harvard University.

Apotheosis of Democracy image by nojustice from Getty Images Signature

> “For the widespread use of machine learning to support the flourishing of democracy, we must be ambitious and imaginative about how we govern predictive tools.”

INTRODUCTION

Machine learning is everywhere. On social media platforms and news sites, in hiring, advertising, mortgage lending, criminal justice, education, and countless other sectors, more and more decisions are being made using predictions generated by algorithms that use complex data processing techniques. AI-evangelists promise that data-driven decision-making will not only boost organizational efficiency, but will also help make organizations fairer and advance social justice. By reducing the scope for human prejudice, irrationality, and error, they claim, machine learning can ensure decisions are made with complete consistency, treating each and every person without regard to morally irrelevant differences.

Yet the effects of machine learning on social justice, human rights, and democracy will depend not on the technology itself, but on human choices about how to design and deploy it. Building and integrating machine learning models into decision-making systems involves choices that prioritize among the interests of different social groups and bake in different fundamental values. Among the most important is whether systems reproduce and entrench pervasive patterns of inequality and how to ensure they do not. How organizations respond to that issue will shape the implications of machine learning for equality, liberty, and fairness, the foundational principles of a flourishing constitutional democracy.

As calls for technology regulation grow stronger and legislators develop concrete proposals, we must keep two important questions at the front of our minds. First, how exactly do machine learning algorithms, and the decision-making systems within which they are used, entrench systemic inequalities? Second, what kinds of concrete goals should governments pursue as they regulate organizations that design and use machine learning to prevent this from happening? This paper sketches answers to both these questions.

We begin from the fundamental starting point that technologies are political, a position developed and defended in its broadest form by Langdon Winner in 1980.[1] A wave of recent scholarship and journalism has provided compelling evidence of how and in what sense machine learning is political, including a book one of us wrote, *Algorithms for the People: Democracy in the Age of AI*, that will be be published by Princeton University Press in fall 2022. In *Algorithms of Oppression*, Safia Noble documents in vivid detail how Google’s search ranks information in ways that often replicate and deepen users’ implicit racial and gender prejudices.[2] Or when algorithms are used to screen prospective new hires, they appear to discount and throw out resumes of women or racial minorities, replicating and amplifying past inequalities in hiring processes.[3] Even though computer scientists create machine learning algorithms that ignore protected characteristics like gender and race, a host of other features tend to be correlated with protected characteristics (called proxies); these features include many outcomes algorithms are asked to predict, thereby causing algorithms to tend to replicate underlying patterns of inequality. Even technologies that appear to embody formal fairness can inadvertently replicate patterns of injustice.[4] Machine learning algorithms can detect, reproduce, and supercharge patterns of discrimination, inequality, and injustice even when they are supposedly agnostic to protected categories. Consider the following illustration.

Imagine for a moment that I run a social media platform that earns revenue through selling and delivering advertisements to users

[1] Langdon Winner, “Do Artifacts Have Politics?” *Daedalus* 109, no. 1 (1980): 121–36.

[2] Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: New York University Press, 2018).

[3] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy, “Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona Spain: ACM, 2020), 15, https://doi.org/10.1145/3351095.3372828.

[4] Cynthia Dwork, Christina Ilvento, and Meena Jagadeesan, “Individual Fairness in Pipelines,” *ArXiv.Org*, 2020; Cynthia Dwork and Christina Ilvento, “Fairness Under Composition,” *ArXiv.Org*, 2018.

who will be interested in them. This is a common industry practice, so I find a team of talented software engineers and data scientists to develop a powerful machine learning system that predicts which ads users will be interested in based on the probability they will click on it. The system is trained on data about which users of the platform tend to click on which kinds of ads. After intensive work to improve the model, we build a system that accurately connects different kinds of ads with specific characteristics of users' profiles and engagement patterns. Suppose I discover that when it comes to job recruitment advertisements on my social network, there are gendered patterns in the kinds of job ads men and women tend to click on. Women are more likely to click on shorter-term service sector or administrative job ads, while men click more on ads for longer-term blue-collar jobs. What's more, the average income attached to the job ads that women tend to click on is significantly lower than the average income for the job ads that men tend to click on. These patterns are not created by the social media site, but the ad targeting system picks up on and replicates them. Unless I intervene, women will be shown ads for jobs with lower average income, entrenching gender inequalities in the workforce and compounding the gender pay gap. Because the system responds to real-time feedback, predicting click probability on the basis of prior clicks, these patterns are replicated and entrenched in an ongoing feedback loop.

Policy makers should be ambitious about how regulations address cases like this. Companies and government agencies must be incentivized and sometimes required to ensure the systems they build don't exacerbate underlying inequalities, and in some cases, instead actively reduce them. In fact, we will argue that in some, if not all cases, there may be no neutral option. If there is no way to build a tool that avoids compounding inequality, organizations must choose to either deliberately build systems that reduce inequalities or accept that their systems will reproduce and entrench them. To ensure the increasingly widespread use of predictive analytics strengthens democracy, we will argue, governments should embed the pursuit of a goal that has previously been explicit in the governance of decision-making and technology design: political equality. We begin by exploring the deficiencies in our current approach to discrimination law. We then define what political equality is and how it would transform our regulatory approach to ensuring machine learning works for democracy.

THE PRINCIPLES UNDERPINNING NON-DISCRIMINATION

The concept policymakers and lawyers usually invoke when thinking about the impact of AI on civil rights is non-discrimination. We believe the use of predictive tools in decision-making will bring to the fore some fundamental limits to, and tensions within, the concept and law of discrimination—tensions that political equality can help illuminate and address. To illustrate this, consider the two principles that underpin current U.S. non-discrimination law: anti-classification and anti-subordination.[5]

Anti-classification embodies a formalistic approach to the principle of equal treatment. According to the anti-classification principle, individual membership in protected groups is morally irrelevant to decisions about the allocation of benefits and burdens: the terms of a mortgage, the success of a job application, and whether someone is granted bail or receives an ad. For that reason, discrimination law prohibits the use of protected traits in decision-making.[6]

Anti-subordination embodies a more substantive approach to the principle of equal treatment. According to the anti-subordination principle, discrimination law aims to eliminate systematic and historical exercises of power of one social group over another, embedded within and entrenched by important decision-making systems, to confront and eradicate relations of subordination and domination. This exercise of power need not be intentional or conscious, though sometimes it will be. Certain social groups are "protected" not because membership in those groups is intrinsically morally irrelevant to decision-making, but because protecting those groups ensures that decision-making systems do not reproduce and exacerbate unjust structures of discrimination and subordination.

The relationship between anti-classification and anti-subordination depends on the case. Let's explore three kinds: in the first, the principles support the same conclusion; in the second, the principles can be stretched to support the same conclusion, but they are often in tension; in the third, the principles are in flat-out contradiction. The progression through these cases tracks the development of the kinds of cases discrimination law has confronted—a stylized history of discrimination.[7]

The first kind of case is straightforward: shop signs that ban black

[5] Victoria F. Nourse and Jane S. Schacter, "The Politics of Legislative Drafting: A Congressional Case Study," *New York University Law Review* 77, no. 3 (2002): 575–624; Richard L. Hasen, "Vote Buying," *California Law Review* 88, no. 5 (2000): 1323–71; David A. Strauss, *The Living Constitution*, Inalienable Rights Series (Oxford: Oxford University Press, 2010).

[6] Jack M. Balkin and Reva B. Siegel, "The American Civil Rights Tradition: Anticlassification or Antisubordination," *Issues in Legal Scholarship* 2, no. 1 (2003): 9–33; David A. Strauss, "Discriminatory Intent and the Taming of Brown," *University of Chicago Law Review* 56, no. 3 (1989): 935–1015; Pamela L. Perry, "Two Faces of Disparate Impact Discrimination," *Fordham Law Review* 59, no. 4 (1991): 523–95; Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104 (2016): 671–732. The UK Supreme Court has described the distinct purposes of prohibitions against direct and indirect discrimination: "The rule against direct discrimination aims to achieve *formal equality of treatment*: there must be no less favourable treatment between otherwise similarly situated people on grounds of colour, race, nationality or ethnic or national origins. Indirect discrimination looks *beyond formal equality* towards a more *substantive equality of results*: criteria which appear neutral on their face may have a disproportionately adverse impact upon people of a particular colour, race, nationality or ethnic or national origins. Direct and indirect discrimination are mutually exclusive.' R v JFS, at 57. Other reasons have been proffered for what motivates the set of prohibited grounds, such as the social meaning and perceived divisiveness of classifications based on race. See Benjamin Eidelson, "Respect, Individualism, and Colorblindness," *Yale Law Journal* 129, no. 6 (2020): 1600–1675; Reva B. Siegel, "From Colorblindness to Antibalkanization: An Emerging Ground of Decision in Race Equality Cases," *Yale Law Journal* 120, no. 6 (2011): 1278–1366; Sophia Moreau, "What Is Discrimination?," *Philosophy & Public Affairs* 38, no. 2 (2010): 143–79.

[7] Owen M. Fiss, "Groups and the Equal Protection Clause," *Philosophy & Public Affairs* 5, no 2 (1976): 171.

people or job ads that ban women violate both anti-classification and anti-subordination principles. Signs that ban black people from the use of public facilities both use a morally irrelevant trait in the distribution of benefits and burdens, and entrench racial domination. While we consign such cases to the dustbins of history, we must not forget that the deliberate exclusion of some groups from public life is the form discrimination has taken for most of human history.

The second kind of case begins to bring out the tension between the principles of anti-classification and anti-subordination. This kind of case historically involved assessing, for instance, whether hiring processes that used factors like education or literacy were legitimate criteria for distinguishing between people for some justified purpose, or whether they were simply a new face on the same old prejudices expressed in public signs and job ads. On the anti-classification view, whether these factors are legitimate should depend on whether they are being intentionally used as proxies for racial and gender categories or are justifiable bases on which to distinguish between applicants for a job. On the anti-subordination view, by contrast, whether these factors are legitimate should depend on the effects of using them in particular decision-making processes on relations of power between citizens. These two modes of reasoning may support the same conclusion, but they may not.

In the third kind of case, the two principles support precisely the opposite actions. Consider our social media job ad hypothetical. Suppose you decide you want to ensure your model advances equality of opportunity with respect to gender, because you think it's the right thing to do and because you think it will win you customers among traditionally underserved groups. You sit down with your computer scientist to figure out how to design a machine learning algorithm to generate accurate predictions about the relevance of certain job ads to individuals without replicating underlying gender inequalities in prior advertisement click data. You find the most effective way to do this is to *include* gender as a variable in the training dataset and the model itself. This allows the model to generate predictions in full knowledge of underlying differences, helping to narrow, although not eliminate, disparities. The best way to avoid replicating historic inequalities in job advertisements, it turns out, is to use gender in the design of your machine learning model.

After you consult your lawyers, however, they explain this action is prohibited by U.S. anti-discrimination law. It involves deliberately using gender to determine who sees which advertisements, violating the anti-classification principle. Here the principle of anti-subordination supports a design choice that violates the principle of anti-classification. When designing and using machine learning models, narrowing outcome disparities across protected groups often requires the explicit use of protected characteristics, an action prohibited by anti-classification. In this kind of case, anti-classification demands exactly the opposite course of action to anti-subordination.[8]

This sharpens the tensions between basing our moral evaluation of decision-making on the legitimacy of particular criteria and basing it on the effects of decisions on relations of power between citizens. On the anti-subordination view, the reason gender is a protected category is that decision-making structures have excluded women and LGBTQ individuals from important opportunities and imposed undue burdens on them for as long as America has existed. Whereas anti-classification asks organizations to hide the complex correlations that characterize our social world, requiring decision-making systems to be designed as if we lived in a color-blind, gender-blind society, anti-subordination requires decision-making systems to be designed in full knowledge of the society in which we actually live.[9]

Over the past half century, anti-classification has come to dominate our understanding of discrimination and the protections of anti-discrimination law. Slowly but surely, courts have narrowed the conditions under which affirmative action is permitted and widened the range of permissible procedures that adversely

[8] Jon Kleinberg et al., "Algorithmic Fairness," *AEA Papers and Proceedings* 108 (2018): 22–27; Pauline T. Kim, "Data-Driven Discrimination at Work," *William and Mary Law Review* 58, no. 3 (2017): 904; Barocas and Selbst, "Big Data's Disparate Impact."

[9] Benjamin Eidelson, *Discrimination and Disrespect*, chap. 2 (Oxford: Oxford University Press, 2015); Deborah Hellman and Sophia Moreau, ed., *Philosophical Foundations of Discrimination Law* (Oxford: Oxford University Press, 2013); Fiss, "Groups and the Equal Protection Clause"; Cynthia Dwork et al., "Fairness through Awareness," *arXiv*: 1104.3913; Jon Kleinberg and Sendhil Mullainathan, "Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability," *arXiv*: 1809.04578.

affect disadvantaged groups.[10] This widening of permissible actions that entrench subordination, along with the failure to comprehensively justify affirmative action, suggests that unless we draw attention to the conflict between these principles, anti-classification may slowly suffocate anti-subordination. Unless the idea of discrimination can be extended beyond the principle of anti-classification, discrimination law risks becoming an increasingly blunt tool for the pursuit of social justice.[11]

Far too often, institutions are prevented from discriminating on the basis of race or gender even when doing so promotes equality. In our example, as you build your social media site, you would not just have no incentive to ensure your machine learning models advance gender equality: even if you wanted to, the law may prevent you from doing so.

> **“In a democracy, which similarities and differences are morally salient is not a settled question to which there is a right answer, but a constant subject of political debate and contest.”**

Machine learning may bring this struggle to a head. The ever more widespread use of machine learning may force a confrontation between the idea that discrimination is wrong because it involves using morally irrelevant criteria in decision-making and the idea that discrimination is wrong because it compounds unjust structures of power.[12] In human decision-making, the tension between these ideas could be overlooked, buried within the opacity of the human mind. We never had to work out what it meant to say a person made a decision because of race or gender, because it is difficult to peer into a person’s mind. The practical constraints on detecting discrimination shielded us from having to work out what makes discrimination wrong. Machine learning may force us to consider how far the idea of discrimination captures what is wrong with using decision-making systems that use legitimate criteria but nonetheless replicate and entrench patterns of social inequality.[13]

This paper sketches an alternative approach to regulating the design and use of machine learning models guided by the ideal of political equality. This would refocus our attention away from the imperative not to discriminate towards the imperative for each and every organization to help secure and promote the conditions necessary to support a healthy democracy.

WHAT IS POLITICAL EQUALITY?

The concept of political equality is fundamental to all political regimes, but especially to democracies. Aristotle’s sharp descriptions of the concept in *Nicomachean Ethics* and *Politics* remain among the most illuminating. Aristotle begins by describing the general principle of equal treatment: like cases should be treated similarly and unlike cases dissimilarly, and more ambitiously,

[10] This narrowing has been less pronounced in the UK. Justice Lady Hale writes, “it is instructive to go through the various iterations of the indirect discrimination concept because it is inconceivable that the later versions were seeking to cut down or to restrict it in ways which the earlier ones did not. The whole trend of equality legislation since it began in the 1970s has been to reinforce the protected given to the principle of equal treatment.” “[T]he prohibition of direct discrimination aims to achieve equality of treatment. Indirect discrimination assumes equality of treatment – the PCP is applied indiscriminately to all – but aims to achieve a level playing field, where people sharing a particular protected characteristic are not subjected to requirements which many of them cannot meet but which cannot be shown to be justified. The prohibition of indirect discrimination thus aims to achieve equality of results in the absence of such justification. It is dealing with hidden barriers which are not easy to anticipate or to spot.” *Essop and others v Home Office*, Judgment, at 10.

[11] Several scholars made this point in the late 1980s and early 1990s. See, for example, Stephen Guest and Alan Milne, eds., *Equality and Discrimination: Essays in Freedom and Justice* (London: University College London, 1985); Iris Marion Young, *Justice and the Politics of Difference* (Princeton, N.J.: Princeton University Press, 1990), 194–98; Christopher McCrudden, “Institutional Discrimination,” *Oxford Journal of Legal Studies* 2, no. 3 (1982): 303–67.

[12] Lily Hu, “What Is ‘Race’ in Algorithmic Discrimination on the Basis of Race?,” *Journal of Moral Philosophy*, Forthcoming.

[13] George Rutherglen, “Concrete or Abstract Conceptions of Discrimination?,” in *Philosophical Foundations of Discrimination Law*, ed. Deborah Hellman and Sophia Moreau (Oxford: Oxford University Press, 2013), 115–37; Eidelson, *Discrimination and Disrespect*; Hugh Collins and Tarunabh Khaitan, *Foundations of Indirect Discrimination Law* (Oxford: Hart Publishing, 2018); Eidelson, “Respect, Individualism, and Colorblindness.”

unlike cases should be treated "in proportion to their unlikeness."[14] This principle is abstract, a formal relationship devoid of substantive content. Part of what makes democracy a distinctive political regime is that citizens argue, in public, about who is similar and different to whom and about the bases on which people should be treated equally. In a democracy, which similarities and differences are morally salient is not a settled question to which there is a right answer, but a constant subject of political debate and contest.[15]

Political equality applies the principle of equal treatment to the allocation of political power. It demands that all citizens are able to participate and engage in public life as equals—that each has the agency to act as an equal citizen. Political equality motivates democratic habits and norms: looking your fellow citizen in the eye regardless of status or wealth or race, opening yourself to others' experiences regardless of how they differ to your own. [16]Different kinds of institutions have embodied the ideal of political equality in democracies at different times. In Aristotle's time, political equality was embodied in the selection of officeholders by lottery; all citizens—which excluded women, foreigners, tradespeople, slaves, and children—were considered capable of rule, so rulers were chosen at random from the entire citizenry. In modern democracy, political equality is embodied in the principle that each citizen's vote counts for the same, no matter how educated or wealthy they are. For much of the history of democracy, the ideal of political equality has motivated reform and revolution, inviting the constant reimagining of social, economic, and political institutions to better approximate its promise. [17]

More recently, Danielle Allen has developed a compelling account of political equality. Allen argues that recent political philosophy has undervalued positive liberties, granting a misguided priority to an individual's negative liberties in the form of rights. Political equality clarifies that negative rights are not prior to, or more fundamental than, positive rights, but that each supports the other. A right to association is not merely a negative right to associate without government interference, it is a positive right to gather with fellow citizens to protect your collective political power and hold your government to account. In the U.S. Bill of Rights, "the right to assemble was closely conjoined to the right to petition political authorities for changes in policies," while today, "the Chinese government" imposes "great restrictions on the freedom of association" not just "to limit freedom of conscience but also to minimize the likelihood that political solidarities will form capable of challenging its authority." [18]

There is widespread understanding of the need to protect the most basic kinds of political equality, for instance in the domain of elections. Policies that seek to secure and protect equal ballot access, for example, demonstrate the value placed on political equality. This kind of basic, formal political equality is relatively uncontroversial, even among more libertarian democratic theorists who prize negative liberties and individual sovereignty above all else. [19]

Allen's notion of political equality extends the concept. It holds that healthy democratic institutions constantly strive to ensure all citizens have equal access to exert civic power and influence. Allen proposes we extend ourselves to see how social and economic injustices can, over time, challenge this kind of political equality even when the law formally secures equal access to civic institutions like the ballot. To illustrate this, consider again our job advertisement example.

Patterns of gender-based difference in job advertisement click rates can burden individuals in many significant ways. For instance, discrimination in access to job information can limit fair equality of opportunity, increase (or at best, leave unchanged) gender wage gaps, and cause unequal access to fundamental workplace-related services like healthcare. Suppose that in the city where you are building your business, these and other foreseeable burdens of discriminatory access to job information compound each other, stratifying the city into the advantaged and the disadvantaged. Those subject to ongoing patterns of discrimination might plausibly face a connected set of barriers to participation in public life, producing lower rates of voting and less time spent engaging in public debate, political activity and, because they lack voice and influence, local decisions that run against their interests. In this case, gender discrimination reproduces and deepens political inequality, blocking some citizens from participating and governing themselves as equals. As you build a machine learning algorithm, the imperative of political equality demands not only that you ensure it avoids compounding underlying inequalities, but that you undertake reasonable efforts to build a system that actively reduces them. That is what's required to live together in a city in which all citizens participate in public life as

[14] A few hundred years earlier, an Aesop's fable told of a fox who invites a crane for dinner, then serves soup in a shallow dish. The fox overlooks a relevant difference; the crane has a long beak, which requires differential treatment, so they need different vessels to drink from. The crane makes the point by inviting the fox for dinner and serving soup in a long, narrow jar. Aristotle, *Nicomachean Ethics*, trans. Roger Crisp (Cambridge: Cambridge University Press, 2000), bk. V, 1131a-b; Danielle S. Allen, *The World of Prometheus: The Politics of Punishing in Democratic Athens* (Princeton, NJ: Princeton University Press, 2000), chap. 11; Frederick Schauer, "On Treating Unlike Cases Alike," in *Symposium on Settled versus Right: A Theory of Precedent* (Minneapolis: University of Minnesota Law School, 2018).

[15] M. S. Lane, *The Birth of Politics: Eight Greek and Roman Political Ideas and Why They Matter* (Princeton, NJ: Princeton University Press, 2014).

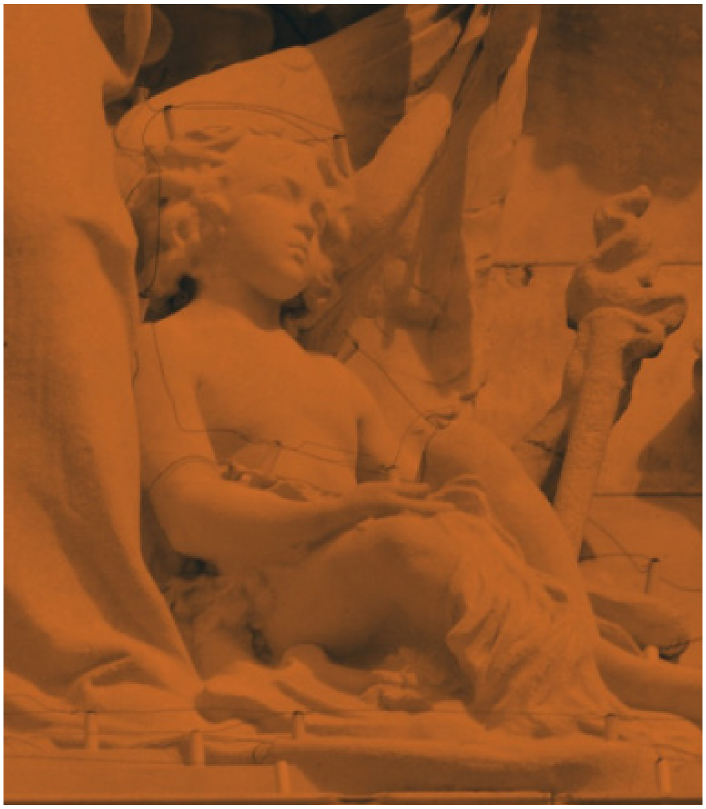
[16] Danielle S. Allen, *Talking to Strangers: Anxieties of Citizenship since Brown v. Board of Education* (Chicago: University of Chicago Press, 2004).

[17] David Runciman, *How Democracy Ends* (New York: Basic Books, 2018); Lane, *The Birth of Politics*.

[18] Danielle S. Allen, "A New Theory of Justice: Difference without Domination," in *Difference without Domination: Pursuing Justice in Diverse Democracies*, edited by Danielle Allen and Rohini Somanathan (Chicago: University of Chicago Press, 2020), 36.

[19] Arthur Ripstein, "Beyond the Harm Principle." *Philosophy & Public Affairs* 34, no. 3 (2006): 215–45, https://doi.org/10.1111/j.1088-4963.2006.00066.x.

equals.[20] Thus, where our vocabulary for discussing AI in terms of non-discrimination led to roadblocks and contradictions, the concept of political equality charts a regulatory path forward.



POLITICAL EQUALITY IN PRACTICE

If we wish to ensure that widespread use of machine learning advances equality among citizens rather than entrenches inequality, our job advertisements example sharpens the kinds of design choices that must be made at the micro level. Organizations must at minimum be permitted, and in some cases required, to use protected categories that serve as proxies for disadvantage to build machine learning systems that empower protected groups. Regardless of what, in technical terms, proves to be the most effective method of building machine learning systems to advance equality, those who design those systems must have clear incentives to build systems that advance equality, including, if necessary, by using protected categories. The question is how to establish laws and regulations, and create public bodies to enforce them, that in practice incentivize or require AI designers and managers to build machine learning systems that advance equality. In our example, what macro structures of law and regulation would provide the incentives or requirements for you to build a job listings recommendation system that advances gender equality in your city?

Political equality explains why we must firmly prioritize the principle of anti-subordination in answering this question. If democracy requires a certain kind of political equality among citizens, then where inequalities between citizens threaten to become entrenched structures of domination or subordination, it may be justifiable, and even necessary, to treat citizens differently to address those inequalities. Inequalities in fundamental goods like access to good jobs or secure housing cannot be allowed to translate into systematic barriers to the capacity of some groups to participate in public life as equals. When these inequalities threaten to become ossified structures of power, political equality provides a language and justification that is rooted in democracy for treating groups differently. Political equality explains when and why we should choose anti-subordination over anti-classification: to protect and secure the kind of equality that democracy requires.

Civil rights and equality laws that were explicitly grounded in the protection of political equality would provide us with a clear principle to guide AI design and governance: we must evaluate how changes in the design and use of machine learning systems impact systemic patterns of inequality that threaten to erode political equality. This principle does not require that those who build machine learning systems must in all cases ensure those systems reduce all instances of inequality. It merely requires that when an inequality is clearly linked to differences in political agency and patterns of subordination, we ensure the decision-making system does not entrench those patterns even if doing so requires violating the principle of anti-classification.

Before we discuss how laws and regulations might structure accountability for securing and advancing political equality, we must define how regulators should think about the risk that a model which reproduces patterns of inequality might threaten political equality over time. A few points are worth clarifying.

First, not all social inequalities are immediately relevant to ensuring citizens can function and participate in public life as equals. Unequal access to wedding cakes based on sexual orientation might, for example, be an egregious issue of disparate treatment we ought to resolve, but it does not obviously threaten political equality. Racial or gendered discrimination in housing or job information access, by contrast, might permit and lead to unequal access to economic opportunity and political decision-making, depressed voter turnout, or racial gerrymandering. This directly contributes to dangerous inequalities in the allocation of political agency.

Second, when and why particular patterns of inequality threaten political equality should be a matter of constant debate and contest. Political equality invites a kind of constant vigilance, a willingness to ensure the rules that govern decision-making

[20] Lawrence R. Jacobs and Theda Skocpol, ed., "American Democracy in an Era of Rising Inequality," in *Inequality and American Democracy* (New York: Russell Sage Foundation, 2005), 1; Tommie Shelby, "Integration, Inequality, and Imperatives of Justice: A Review Essay," *Philosophy & Public Affairs* 42, no. 3 (2014): 253–85, https://doi.org/10.1111/papa.12034; J. Phillip Thompson, "Politics in a Racially Segregated Nation," in *The Dream Revisited: Contemporary Debates About Housing, Segregation, and Opportunity*, edited by Ingrid Ellen and Justin Steil (New York/Chichester, West Sussex: Columbia University Press, 2019), 190–93; Elizabeth Anderson, "Five. Democratic Ideals and Segregation," in *The Imperative of Integration* (Princeton, NJ: Princeton University Press, 2015), 89–111.

remove concrete barriers to participation and engagement that some citizens face. It requires a connection, in other words, between the rules and the sociological patterns of the real world within which those rules operate. There is no settled or right answer to the questions of power and agency in democratic life that political equality invites us to wrestle with.

With these considerations in mind, a focus on political equality provides a clear way to determine when we should be concerned about the impact of a particular machine learning model and the decision-making system within which it is deployed. It invites us to focus, first, on the social groups whose power and agency the system might affect, and second, on the role the institution that uses the system plays in shaping the power and agency of those groups. It enables laws and regulations to embrace differentiations between the incentives and requirements imposed on different institutions and across different social groups.

POLITICAL EQUALITY FOR SOCIAL GROUPS

The overriding concern of political equality is that some groups of citizens are not subject to insurmountable and immovable barriers to participation and engagement in public life. It leaves open the question of which are the categories on the basis of which some citizens experience these barriers, whether race or gender, sexual identity, socioeconomic class, or geography.

Political equality would invite concern about any machine learning system that has disparate impact across racial groups, for example. It would justify the use of race as a basis for differential treatment because race is a proxy for centuries of domination and exclusion from practices of reciprocity that is itself differentially experienced. Race is a crude proxy for disadvantage, because the relationship between race and disadvantage is contingent rather than inexorable, and yet because race has been among the most persistent categories for treating people differently in American history, it is also a pervasive proxy for disadvantage.

Contrast race with another barrier to political equality: geography, a neglected category of disadvantage.[21] People born in neighborhoods with lower average incomes, less access to capital and investment, and poorer education and healthcare systems are subject to a range of connected decision systems that make it systematically more difficult for them to function as political equals. Insofar as geography is a practical barrier to political equality, then in relevant decision contexts, political equality may treat geography as a legitimate basis for treating people differently. For instance, if geography is driving exclusion, polarization, and stratification in housing, geography may be legitimate criteria to use in decision-making systems that shape access to housing. [22]

Political equality also clarifies that justifications of differential treatment across groups do not flow in both directions. The fact that gender or race is a category of persistent disadvantage justifies positive action on behalf of those who are disadvantaged, not those who are advantaged. The fact that gender is a category of disadvantage justifies positive action not on the grounds of gender, but on behalf of women, because women are subject to the myriad consequences of that disadvantage.

ROLE OF INSTITUTIONS IN POLITICAL EQUALITY

Political equality also supports principled distinctions between the incentives and requirements imposed on different institutions, focusing attention on how institutions affect the capacity of citizens to function as equals. When institutions use machine learning to control access to something fundamental to citizenship, such as freedom from arbitrary treatment by law enforcement, this poses a greater threat to political equality than when businesses use it to do something comparatively trivial, such as provide (or refuse to provide) wedding cakes.

The concept of basic interests is helpful. People have "basic interests in the security, nutrition, health, and education needed to develop into, and live as, a normal adult. This includes developing the capacities needed to function effectively in the prevailing economic, technological, and institutional system, governed as a democracy, over the course of their lives." The more critical a good or service to securing a basic interest, the greater the risk the institution that controls that good or service will cement domination and corrode reciprocity. The greater the threat an institution poses to political equality,

[21] Michael C. Lens, "Measuring the Geography of Opportunity," *Progress in Human Geography* 41, no. 1 (2017): 3–25; Philip McCann, *The UK Regional–National Economic Problem: Geography, Globalisation and Governance* (London: Routledge, 2016); Andy Peter Edward, "The Geography of Inequality: Where and by How Much Has Income Distribution Changed since 1990?" Working Paper 341," *IDEAS Working Paper Series from RePEc*, 2013; Abhijit V. Banerjee and Esther Duflo, *Good Economics for Hard Times* (New York: PublicAffairs, 2019); Benjamin Austin, Edward Glaeser, and Lawrence Summers, "Jobs for the Heartland: Place-Based Policies in 21st-Century America," *Brookings Papers on Economic Activity*, Spring 2018, 151–255. Susan Sturn offers a compelling account of how to reframe affirmative action in education by "(1) nesting it within an effort to transform institutions to ensure full participation, (2) shifting from rewarding privilege to cultivating potential and increasing mobility, and (3) building partnerships and enabling systemic approaches to increasing educational access and success...these structural approaches are less likely to trigger strict scrutiny from the courts, and will foster the inquiry needed to document the need for affirmative action in admissions and expand the justifications for race-conscious approaches." Susan P. Sturm, "Reframing Affirmative Action: From Diversity to Mobility and Full Participation," *University of Chicago Law Review Online*, October 30, 2020, https://lawreviewblog.uchicago.edu/2020/10/30/aa-sturm/.

[22] Danielle Allen, "Talent Is Everywhere: Using ZIP Codes and Merit to Enhance Diversity," in *The Future of Affirmative Action: New Paths to Higher Education Diversity after* Fisher v. University of Texas, ed. Richard D. Kahlenberg (New York: Century Foundation Press, 2014), 151; Michael J. Sandel, *The Tyranny of Merit: What's Become of the Common Good?* (New York: Farrar, Straus and Giroux, 2020).

the more stringent the obligations imposed on it should be.[23]

For instance, institutions that provide access to housing may fundamentally shape the capacity of citizens to function as equals. This would include mortgage providers and other lenders, but it may also include organizations that match vacancies in housing markets to people who might be interested in those vacancies or the advertising systems of social media companies that show different ads for mortgages or houses to different people. By contrast, websites that merely recommended different kinds of home furnishings might not be subject to the same kinds of requirements.

An institution's role in securing citizens' basic interests contrasts with the more common focus on whether an institution is a public body or private company. Many goods and services necessary for citizens to function as equals are provided by private companies, and political equality invites us to consider structuring incentives and imposing requirements on those companies which ensure the machine learning systems they build secure and protect political equality among citizens over time. Political equality roots the obligations imposed on institutions not in their legal status, but in their role in securing the conditions of political equality over time.[24]

REFORMING CIVIL RIGHTS AND EQUALITY LAW

The idea of political equality suggests two important reforms to the macro structure of how we regulate decision-making: the first to do with how we structure requirements that shape how organizations build and use machine learning, the second to do with how we monitor and enforce those requirements.

Positive Equality Duties

At present, the primary duty that civil rights and equality law rely on is the duty not to discriminate. We have already explored the tensions that underpin this duty. We propose that the duty not to discriminate should be more narrowly targeted, focused on what U.S. discrimination calls "disparate treatment" and what UK and EU discrimination law calls "direct discrimination," and should be subsumed under a wider category of Positive Equality Duties (PEDs).

PEDs would require government agencies and private companies in defined sectors and contexts to demonstrate that they have taken reasonable steps to consider how best to advance equality among protected and non-protected groups. This would require institutions to take preemptive measures to evaluate the impact of decision-making systems, compare alternative ways of designing them, and take reasonable measures to understand and address disparities across protected and non-protected groups. There would be a legal presumption that when protected characteristics are used as part of reasonable efforts to discharge a PED, and there is a strong basis in evidence that doing so will reduce inequalities across protected groups, the use of protected traits will not violate non-discrimination law. PEDs would permit organizations to treat different people differently for the purpose of addressing concentrated disadvantage, "based on the recognition that equal treatment...may lead to an unequal outcome, and that therefore preferential treatment is needed."[25] Like the Constitution of South Africa, deliberately written to confront the country's violent history of racial oppression, we should understand PEDs not as "a deviation from, or invasive of, the right to equality," not "'reverse discrimination' or 'positive discrimination,'" but rather, as "integral to the reach of our equality protection."[26]

While PEDs would represent a stark but necessary shift in how regulators govern the design and use of machine learning models, they are not wholly unprecedented. In many industries, regulators require companies to work proactively to preempt possible negative consequences of their products. In the U.S., cigarette companies are required to post health warnings on their products. In the EU, they must package cigarettes with alarmingly graphic depictions of smoking-induced medical conditions. These packaging and marketing requirements might reasonably be rephrased as "positive health duties." We can understand them as the requirements regulators have imposed after deciding to prioritize public health as a social good over the profits and absolute freedom of cigarette manufacturers. Similarly, environmental requirements imposed on auto manufacturers might be reinterpreted as "positive environmental duties," as regulators imposed these requirements after recognizing the importance of clean air as a social good. Because political equality underpins a secure and stable democracy,

[23] Ian Shapiro, "On Non-Domination," *University of Toronto Law Journal* 62, no. 3 (2012): 294; Ian Shapiro, *Politics against Domination* (Cambridge, MA: Belknap Press, 2016); Ian Shapiro, *Democratic Justice* (New Haven, CT: Yale University Press, 1999).

[24] Chiara Cordelli, *The Privatized State* (Princeton, NJ: Princeton University Press, 2020); Virginia Eubanks, "A Child Abuse Prediction Model Fails Poor Families," *Wired*, January 15, 2018, https://www.wired.com/story/excerpt-from-automating-inequality/.

[25] Christa Tobler, "Limits and Potential of the Concept of Indirect Discrimination," Report, European Commission, 2008, 51.

[26] Aileen McColgan, *Discrimination, Equality and the Law*, Human Rights Law in Perspective (London: Hart Publishing/Bloomsbury Publishing Plc, 2014), 8–9, chap. 3, quote at 78; J Ackermann, *National Coalition for Gay and Lesbian Equality v Minister of Justice* (SA CC 1998); Sandra Fredman, "Addressing Disparate Impact: Indirect Discrimination and the Public Sector Equality Duty," *Industrial Law Journal (London)* 43, no. 3 (2014): 349–63. PEDs could be modeled on the UK Equality Act's provisions for deliberately advancing equality, although it would significantly extend them. Until the Equality Act of 2010, UK law approached positive action, positive duties, and other measures explicitly designed to promote substantive equality as exceptions to the "general principle of non-discrimination." "UK law [did] not permit 'reverse discrimination" other than for narrowly defined purposes, such as "positive measures to afford access to training and to encourage under-represented groups to take up employment." The EA then established the Public Sector Equality Duty (PSED) that applied to public bodies and non-public bodies performing public functions in relation to those functions. The PSED requires such bodies to give "due regard" to a number of statutory needs, including the need to eliminate unlawful discrimination, advance equality of opportunity, and foster good relations between persons defined by reference to protected characteristics. No matter which functions the relevant bodies are performing, adequate consideration must be given to equality defined in terms of these statutory needs. Catherine Barnard and Bob Hepple, "Substantive Equality," *Cambridge Law Journal* 59, no. 3 (2000): 576.

regulators should require machine learning designers to fulfill specific positive equality duties to strengthen our democracy.

PEDs would transform the governance of decision-making. They would require institutions to directly confront disadvantages that follow from membership in protected groups, and more broadly, to undertake measures to encourage participation in public life by those groups. As the Clinton administration's Affirmative Action Review put it, PEDs would require institutions "to expand opportunity for women or racial, ethnic, and national origin minorities by using membership in those groups that have been subject to discrimination."[27] As the scholar Virginia Eubanks argued, predictive "tools … left on their own, will produce towering inequalities unless" they are "built to explicitly dismantle structural inequalities, their increased speed and vast scale [will] intensify them dramatically."[28] Given this, positive duties may be "the most appropriate way to advance equality and to fight discrimination, including indirect discrimination."[29]

Our purpose is not to define the precise content of PEDs, as they should vary considerably across different institutions and with respect to different social groups, and of course, across legal systems. Instead we want to explore a more pressing proposal that relates to how PEDs would be developed and enforced by regulators as social conditions change and technology evolves.

A New Role for Regulators

Political equality invites a rethink in not only the content of duties to advance equality, but also in how those duties are monitored and enforced. The shift we've proposed away from negative prohibitions against discrimination towards positive duties to advance equality should be accompanied by a shift from ex post evaluation of decision-making systems by courts, to an ex ante evaluation of decision-making systems by those who design and use them, overseen by regulators tasked with monitoring and enforcing equality duties.

> "As you build a machine learning algorithm, the imperative of political equality demands not only that you ensure it avoids compounding underlying inequalities, but that you undertake reasonable efforts to build a system that actively reduces them."

Think of machine learning models integrated into decision-making systems as a kind of infrastructure, a connected set of systems that are built into our social, economic, and political environment and forgotten about until we come to update them. In infrastructural contexts, organizations often have ex ante duties to surface and evaluate harms that might be caused by different ways of building that infrastructure, whether through environmental or other kinds of impact assessments. We should think of enforcing equality duties in a similar fashion. Instead of defining precisely how organizations should build and integrate machine learning models, or attempting to evaluate individual systems before they are built, regulators should establish broad duties for organizations to evaluate systems as they are designed and while they are deployed. [30]

This could be achieved either by granting new powers to existing civil rights regulators such as the US Department of Housing

[27] George Stephanopoulos and Christopher F. Edley, *Affirmative Action Review: Report to the President* (Washington, DC: White House, 1995; John Valery White, "What Is Affirmative Action?," *Tulane Law Review* 78 (2004): 2117–2329.

[28] Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's Press, 2017), chap. 5.

[29] Tobler, "Limits and Potential of the Concept of Indirect Discrimination," 52.

[30] Raghavan et al., "Mitigating Bias in Algorithmic Hiring," 469–81; Simonetta Manfredi, Lucy Vickers, and Kate Clayton-Hathway, "The Public Sector Equality Duty: Enforcing Equality Rights through Second-Generation Regulation," *Industrial Law Journal* 47, no. 3 (2018): 365–98.

and Urban Development, the US Department of Labor, and the various discrimination and consumer protection programs of the US Federal Trade Commission, or by establishing an entirely new regulator tasked with regulating the design and use of machine learning and AI systems, such as an AI Platform Agency (APA). Such an agency could deploy several constitutionally permitted methods for establishing and monitoring positive equality duties. They could simply establish incentives for organizations to evaluate the impact of machine learning models on protected groups before they are deployed and mitigate any adverse effects, for instance through tax breaks or the risk of substantial fines. Such an agency could also periodically request to see Equality Impact Assessments (EIAs) that organizations would be required to complete in advance of deploying machine learning models and even to request access to certain datasets to verify the information contained in EIAs.[31]

Any regulator should develop a range of tools to monitor the enforcement of PEDs and those tools should be flexibly applied across different sectors and organizations, and with respect to different groups. For instance, large and well-resourced private companies whose machine learning systems have profound impacts on relations of equality over time, such as mortgage lenders, credit agencies, or technology companies like Facebook and Google, might be legally required to complete EIAs for all major systems they design and deploy, and to periodically submit those reports to relevant regulators. By contrast, small, under-resourced companies such as the start-up social media platform we examined might have less burdensome requirements, requiring only that they complete an annual equality impact evaluation without the need to submit it to regulators. PEDs importantly establish "reasonable duties" to evaluate and address equality impacts, and the reasonableness criteria would take account of organizational capacity and the relevant decision-making arena. [32]

CONCLUSION

This kind of approach is exactly the kind of governance regime that predictive tools like machine learning make possible—and necessary. Realizing the ambition of President Obama's report will require something very like the approach we have described, informed by political equality: "To avoid exacerbating biases by encoding them into technological systems, we need to develop a principle of 'equal opportunity by design'—designing data systems that promote fairness and safeguard against discrimination from the first step of the engineering process and continuing throughout their lifespan." [33]

Political equality supports an approach in which private companies and public bodies routinely record, report, and justify disparities in outcomes produced by predictive tools. This approach would institutionalize the asking of exactly those questions that political equality invites and that discrimination encourages us to ignore. For the widespread use of machine learning to support the flourishing of democracy, we must be ambitious and imaginative about how we govern predictive tools. Positive Equality Duties and an AI Platform Agency offer a vision of how we might begin to do that.

[31] Michael Veale and Reuben Binns, "Fairer Machine Learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data," *Big Data & Society* 4, no. 2 (2017), https://doi.org/10.1177/2053951717743530. Veale and Binns suggest that for smaller firms, third parties could hold data to help with contextual evaluations of fairness.

[32] Swee Leng Harris, "Data Protection Impact Assessments as Rule of Law Governance Mechanisms," *Data & Policy* 2 (2020); IFOW, "Mind the Gap: The Final Report of the Equality Task Force," 50–54 (Institute for the Future of Work, 2020), accessed February 24, 2022, https://www.ifow.org/publications/mind-the-gap-the-final-report-of-the-equality-task-force.

[33] Executive Office of the President, "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights," May 2016, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

Technology and Democracy Discussion Paper Series

Justice, Health, and Democracy Impact Initiative &
Carr Center for Human Rights Policy
Harvard University
Cambridge, MA 02138

Statements and views expressed in this paper are solely those of the author and do not imply endorsement by Harvard University, the Justice, Health, and Democracy Impact Initiative, or the Carr Center for Human Rights Policy.