Yale Journal of Law & Technology
Volume 27, Issue 1

Fair-Enough AI

Jane R. Bambauer[*] and Tal Z. Zarsky[**]

AI is unfair. It can be inaccurate (in several ways), biased (in several ways, and to several groups), disproportionate, exploitable, and opaque. The policy world is awash in AI-governance frameworks, ethical guidelines, and other policy documents, but these lack concrete standards and provide little guidance on how to select between competing versions of (un)fairness. In other words, they abdicate the responsibility of setting priorities among values. At the same time, many of the policy documents harshly criticize AI and algorithmic tools for deficiencies in some particular aspect of fairness without considering whether alternative designs that fix the problem would make the system more "unfair" in other aspects. Ad-hoc criticism is abundant and hard to predict.

This article offers a strategy for AI ethics officers to navigate the "abdication/ad-hoc criticism" problems in the regulatory landscape. After explaining the meaning and sources of the most important forms of AI unfairness, we argue that AI developers should make the inevitable tradeoffs between fairness goals as consciously and intentionally as the context will allow. Beyond that, in the absence of clear legal requirements to prioritize one form of fairness over another, an algorithm that makes well-considered trade-offs between values should be deemed "fair enough."

* Brechner Eminent Scholar, University of Florida Levin College of Law and College of Journalism and Communications.
** Dean and Professor of Law, University of Haifa Faculty of Law. The authors are grateful for comments and feedback on early drafts from Anthony Bok, John Villasenor and Diana Yakowitz.

Article Contents

Introduction

Suppose a state court system is considering adopting a data-driven scoring tool that will help judges assess whether criminal defendants would pose a risk to the public if released before trial. The court has already ruled out several products from commercial vendors, but debate has fractured over three remaining contenders. Option A makes fairly accurate predictions overall, but is more likely to be wrong (giving an inflated prediction of risk) for Black and male defendants. Option B has no bias in errors across race and gender categories. However, its error rate for *all* demographic categories is slightly higher. Option C is as accurate as Option A and as neutral as Option B. However, to achieve greater accuracy and neutrality, it uses a larger amount of sensitive personal data, thereby imposing a privacy cost to the defendants (and to others whose data may be collected or analyzed in the process of developing the tool).

Which option is the most *fair*? Are they all fair enough? Do the answers change if the alternative—judicial decisions without guidance from a predictive score—tend in the aggregate to be less accurate, more biased, *and* more privacy invasive? And what would be a proper methodology for selecting among them?

State, national, and transnational governing bodies have published several frameworks and blueprints that mark an intent to regulate nearly every ethical implication of AI,[1] but none of them answer the realistic hypothetical posed above. At this point, it is commonplace and even trite to remind readers that fairness is an ambiguous concept. While scholars are in a self-aware struggle to

---

[1] OFF. OF SCI. & TECH. POL'Y, BLUEPRINT FOR AN AI BILL OF RIGHTS: MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE 5-7 (2022), https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf [https://perma.cc/PL94-NAVN] [hereinafter *Blueprint for an AI Bill of Rights*]; Exec. Order No. 14,110, 88 Fed. Reg. 75191, 75191-93 (Nov. 1, 2023); IDAHO CODE § 19-1910 (2019) (imposing transparency requirements on pretrial risk assessment tools); Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828, 2024 O.J. (L 119) 1, 1 [hereinafter EU AI Act] (setting out to "lay[] down a uniform legal framework . . . to promote the uptake of human centric and trustworthy artificial intelligence . . . .").

define AI fairness or, at least, to harmonize its many definitions,[2] regulators are unwittingly setting an AI fairness trap. By and large, the guidance documents lawmakers have released do not explain which forms of fairness should be honored when there are conflicts between them, yet they also exhibit little tolerance for fairness flaws of any kind within the regulation scheme.

On one hand, regulatory plans across Europe and the United States contain promises of abstract and ambitious policy goals such as "transparency," "privacy," "accuracy," "access," and freedom from "bias." [3] Some AI applications might be able to make improvements along one of these social goals without undermining others, but at a certain point, when they have reached the "Pareto frontier,"[4] an AI developer will not be able to advance without reconciling the various fairness goals through tradeoffs. [5] Most

---

[2] *See, e.g.,* MICHAEL KEARNS & AARON ROTH, THE ETHICAL ALGORITHM: THE SCIENCE OF SOCIALLY AWARE ALGORITHM DESIGN 85-86 (2020) (describing tradeoffs between different kinds of fairness); Arvind Narayanan, *Tutorial: 21 Fairness Definitions and Their Politics,* YOUTUBE (Mar. 1, 2018), https://www.youtube.com/watch?v=jIXIuYdnyyk (discussing the "lack of a concrete definition of 'fairness'" at 1:53); Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113, 138-39 (2018); Lindsay Weinberg, *Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches,* 74 J.A.I. RES. 75, 79-80 (2022).

[3] *See supra* note 1; *see also* Assemb. Con. Res. 96, 2023-2024 Assemb., Reg. Sess. (Cal. 2023) (expressing the support of the California legislature for the 23 Asilomar AI Principles, which include "Failure Transparency," "Judicial Transparency," "Personal Privacy," and "Liberty and Privacy").

[4] Shizhou Xu & Thomas Strohmer, *Fair Data Representation for Machine Learning at the Pareto Frontier*, 24 J. MACH. LEARNING RSCH. 1, 6 (2023) (explaining technical issues with computing the Preto frontier, "the optimal trade-off . . . between prediction accuracy and fairness.").

[5] On the privacy-versus-accuracy tradeoff, see Kleinberg et al., *supra* note 2, at 150 ("The existence of disparate impact is clear . . . . Then the question is the standard one: Can the disparate impact be justified, given the relevant standard? That is the same question that would be asked if an algorithm were not involved. The presence of the algorithm goes further—it makes it possible to quantify the tradeoffs that are relevant to determining whether there is 'business necessity' (or some other justification for disparate impact)."). On the privacy-versus-equity tradeoff, see Alice Xiang, *Being 'Seen' Versus 'Mis-Seen'; Tensions Between Privacy and Fairness in Computer Vision*, 36 HARV. J.L. & TECH. 1, 45-49 (2022). On the equity-versus-accuracy tradeoff, see Sam Corbett-Davies, Emma Pierson,

government frameworks do not even recognize the variety of meanings of those concepts, let alone provide a mechanism for mediating between them when there is tension.[6] This is an abdication problem—that is, lawmakers discuss rights and obligations using abstract terms that are neither self-defining nor independent of one another.

On the other hand, these same lawmakers have also chastised companies on the basis that their algorithmic decision-making tools have violated one specific AI ethic or another—by creating models

---

Avi Feller, Sharad Goel, Aziz Huq, *Algorithmic Decision Making and the Cost of Fairness*, PROC. 23D ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING, Aug. 4, 2017, at 797, 802. For equality-versus-equality tradeoffs, see Sandra Mayson, *Bias In, Bias Out,* 128 YALE L.J. 2218, 2249-51 (2019).

[6] For example, the European Union's AI Act requires the protection of privacy, EU AI Act, *supra* note 1, at 58, protection against bias, *id.* at 57, accuracy, *id.* at 61, consistency, *id.*, and comprehensibility, *id.* at 59. Despite its length (144 pages), the Act does not establish minimum requirements, nor does it establish relative priorities between the obligations. The White House Blueprint for an AI Bill of Rights promises protection from biased or inaccurate systems, but also promises rights to an explanation, to a human alternative, and to control how personal data is used. *Blueprint for an AI Bill of Rights*, *supra* note 1 at 5-7. Each specific right is also vague. For example, AI firms are obligated to ensure "equity" which is defined as "the consistent and systematic fair, just, and impartial treatment of all individuals." *Id.* at 10. It goes on to explain that "[s]ystemic, fair, and just treatment must take into account the status of individuals who belong to underserved communities" but does not describe *how* firms should take these factors into account. *Id*. at 10. And California's Consumer Privacy Protection Agency has released draft rules that require users of automated decision-making systems to honor any request to opt out of the system unless they can rebut the presumption that there are alternative means of making the decisions without automation that are valid, feasible, and "fair" enough to work as a substitute. CAL. PRIV. PROT. AGENCY, DRAFT AUTOMATED DECISIONMAKING TECHNOLOGY REGULATIONS 10 (2023), https://cppa.ca.gov/meetings/materials/20231208_item2_draft.pdf [https://perma.cc/2DR8-V46Z]. One exception, however, is the AI Risk Management Framework promulgated by the National Institutes of Standards and Technology (NIST). NAT'L INST. OF STANDARDS & TECH., ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK 12 (2023), https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf [https://perma.cc/JCL9-7HAD] (identifying seven characteristics of trustworthy AI: validity/reliability, safety, security/resilience, explainability, privacy-enhancement, management of bias, and accountability/transparency) [hereinafter NIST AI RISK MANAGEMENT FRAMEWORK].

that are too biased, or too opaque, for example.[7] But these critiques are overdetermined. It is impossible to design a system that has the best possible performance across *all* versions of AI fairness. To be sure, there are algorithmic tools that are needlessly inaccurate, biased, or opaque, and those are proper targets for regulation. But criticism abounds for AI tools even when they may be operating at the point at which improvement along one ethical dimension would degrade performance along another. An algorithm that is as accurate as possible may have more racial bias, and an effort to reduce that bias without reducing accuracy may require the collection of more data (and, thus, a reduction in privacy) or steps to reduce gaming (a reduction in transparency). If priorities between values are not set in advance, a regulator will always be able to find a side of the prism through which the AI looks unfair.

Consider the following guideline from the White House *Blueprint for an AI Bill of Rights* addressing bias and discrimination. [8] The *Blueprint* states that "[a]lgorithmic discrimination occurs when automated systems contribute to unjustified different treatment or impacts disfavoring" members of historically disadvantaged groups, and urges AI developers to run tests simulating real-world contexts to see if the automated system

---

[7] *See infra* notes 8-11 and accompanying text. *Compare* Natasha Lomas, *Elon Musk's X Taken to Court in Ireland for Grabbing EU User Data to Train Grok Without Consent,* TECHCRUNCH (Aug. 7, 2024, 3:02 AM PDT), https://techcrunch.com/2024/08/07/elon-musks-x-taken-to-court-in-ireland-for-grabbing-eu-user-data-to-train-grok-without-consent/ [https://perma.cc/S52M-TWQ4] (describing Ireland's Data Protection Authority's case against X for using EU user data for ongoing chatbot training) *with* EUR. UNION AGENCY FOR FUNDAMENTAL RTS., BIAS IN ALGORITHMS: ARTIFICIAL INTELLIGENCE AND DISCRIMINATION 11-12, 26, 53, 72 (2022), https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf [https://perma.cc/E8N5-C2MU] (discussing bias in algorithms trained to detect offensive speech concluding "assessment of bias in view of their actual use").

[8] *Blueprint for an AI Bill of Rights*, *supra* note 1. The topic of algorithmic discrimination or bias is one of the most pressing concerns for AI ethicists and has been for years. *See generally* CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION (2016) (discussing the algorithmic discrimination and bias in a variety of contexts including college admissions, criminal justice, and employment); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact,* 104 CALIF. L. REV. 671 (2016) (discussing algorithmic discrimination arising from discriminatory data and potential legal reforms to address it in the employment context).

will "produce disparities."[9] However, the *Blueprint* does not specify which disparities should be considered suspect, and when (or why) disparities might be tolerable or well justified. The *Blueprint*'s illustrations drawn from real-world examples do not alleviate ambiguity.

Some of the illustrations are clear because they involve design flaws, such that bias can be reduced without any negative impact on other dimensions of fairness.[10] For example, the *Blueprint* criticizes an algorithm deployed in the medical context that used prior health care costs as a proxy for patient health.[11] However, since Black patients in that dataset tended to decline treatment more often than other patients for a variety of financial and other reasons, past health costs were actually *not* a good measure of health needs. A Black patient who needed the same treatment as a white one but declined that service looked healthier in the historical data used to train the algorithm. As a result, the tool steered doctors to offer fewer medical interventions to Black patients than equally sick white patients.[12] This is an example of a *biased objective function* that may very well be improved without reducing fairness along other dimensions.[13]

Another example concerns an algorithm that more frequently recommended cesarean sections for pregnant Black women because it failed to control for other factors correlated with vaginal-birth success rates, such as marital status and type of insurance.[14] If these additional factors were added into the model, the algorithm may very well have produced more accurate *and* less biased results.[15]

But some disparities, of some sort, for some group, are often

---

[9] *Blueprint for an AI Bill of Rights*, *supra* note 1, at 5, 27.
[10] *Id.* at 25.
[11] *Id.* (discussing Ziad Obermeyer, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations,* 366 SCIENCE 447 (2019)).
[12] Obermeyer et al., *supra* note 11, at 450.
[13] *See infra* Section III.B.
[14] *Blueprint for an AI Bill of Rights*, *supra* note 1 at 25; Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones, *Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms,* 383 N. ENGL. J. MED. 874, 875, 877, 879 (2020) (discussing the results reported in William A. Grobman et al., *Development of a Nomogram for Prediction of Vaginal Birth After Cesarean Delivery*, 109 OBSTETRICS & GYNECOLOGY 806 (2007)).
[15] The effect on privacy would be ambiguous: the hospital would need to collect or repurpose marital status and insurance type, but would not need to ask about race.

unavoidable.[16] The Blueprint provides little direction for socially responsible developers who have already optimized their systems to the point where trade-offs between forms of fairness must be made. As a result, the Blueprint invites unlimited opportunity for criticism.

Consider the very first example of a discriminatory algorithm provided in the report:

> An automated system using nontraditional factors such as educational attainment and employment history as part of its loan underwriting and pricing model was found to be much more likely to charge an applicant who attended a Historically Black College or University (HBCU) higher loan prices for refinancing a student loan than an applicant who did not attend an HBCU. This was found to be true even when controlling for other credit-related factors.[17]

The study cited by the *Blueprint* tested loan offerings from a FinTech company using just three fictional profiles. These test profiles involved otherwise-identical loan applicants who had graduated from Howard University (the HBCU), New Mexico State University (tested as a Hispanic-Serving Institution), or New York University (NYU).[18]

It is hard to understand what to make of the example given the difference between NYU and the other two schools in terms of student ability, costs of education, and the average commercial value of a degree.[19] Would the charge of racial bias still hold if the

---

[16] John Kleinberg, Sendhil Mullainathan &Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, ARXIV 15 (Nov. 17, 2016), https://arxiv.org/pdf/1609.05807 [https://perma.cc/L2ZW-Q5T8]. Thus, in addition to the first order tradeoffs between different types of fairness (accuracy, avoiding bias, privacy, etc.), there are also some tradeoffs that may need to be made *within* a particular form of fairness. Which groups should be protected from bias? And which types of outcomes or errors must be equalized?

[17] *Blueprint for an AI Bill of Rights*, *supra* note 1, at 24.

[18] STUDENT BORROWER PROT. CTR., EDUCATIONAL REDLINING 18 (2020), https://protectborrowers.org/wp-content/uploads/2020/02/Education-Redlining-Report.pdf [https://perma.cc/2VNM-EFW6].

[19] NYU's 25th-percentile SAT score for incoming freshmen is about 200 points higher than the 75th percentile at Howard, and the schools have very different rates of employment at graduation and long-term income as well. *Compare New*

automated system offered better interest rates to NYU graduates, but identical rates to graduates of Howard and the University of Idaho, a predominantly white school with similar entering credentials and career prospects to Howard? Does it matter whether the FinTech company's algorithm produces biased *errors* (i.e., whether Black loan applicants default less often than white loan applicants who are offered the same terms), or are differential recommendation rates alone sufficient to support a charge of bias? Should the FinTech company intentionally constrain its algorithm from considering some factors that correlate with race, such as college attended, even if those factors improve the accuracy? How much reduction in accuracy (and resulting increase in costs and interest rates) is fair to impose on all loan applicants?

None of these questions can be answered from a close reading of the *Blueprint*, and yet its illustrative guidance engages in ex post critiques that assume the substantive goals of AI ethics are obvious and well-settled. This is the abdication/ad-hoc critique problem in action: a lack of substantive benchmarks and priorities upfront, and a bottomless well of criticism and backlash in hindsight. If these frameworks are enforced, they will either become empty, check-the-box procedures that industry can meet by declaring that almost anything is "fair,"[20] or be interpreted ex post, depriving industry of notice and stability. It is a recipe for slow and highly neurotic development in the AI space—and a regulatory structure that itself

---

York University Admissions & Applying Information, U.S. NEWS & WORLD REP., https://www.usnews.com/best-colleges/nyu-2785/applying [https://perma.cc/44XR-L39Q], *with Howard University Admissions*, *U*.S. NEWS & WORLD REP., https://www.usnews.com/best-colleges/howard-university-1448/applying [https://perma.cc/GB9H-XA4D]. Moreover, NYU graduates can take advantage of programs such as the Public Service Loan Forgiveness program that permits graduates to reduce the cost of their existing student-loan debt. *Student Loan Information & Repayment*, N.Y. UNIV., https://www.nyu.edu/students/student-success/financial-education-at-nyu/public-service-loan-forgiveness.html [https://perma.cc/AC95-2K8A].

[20] Laurie Clarke, *AI Auditing Is the Next Big Thing. But Will it Ensure Ethical Algorithms?,* TECHMONITOR (Apr. 14, 2021), https://techmonitor.ai/technology/ai-auditing-next-big-thing-will-it-ensure-ethical-algorithms [https://perma.cc/YJP6-GJYS] ("'There is no consensus on what an audit means,' says Mona Sloane, a senior research scientist at the NYU Centre for Responsible AI and a fellow at the NYU Institute for Public Knowledge. 'We don't even know what 'bias' means or what 'harm' means, so that is a real concern.'")

lacks fairness.

The abdication/ad-hoc critique problem is not purposeful or malicious. Regulators have not set out to trick the AI industry. Rather, it is the entirely natural product of a political and human instinct to avoid making hard tradeoffs explicit, especially when reasonable people disagree on how the tradeoffs should be made.[21] Lawmakers always encounter difficulty when one popular objective must be sacrificed in favor of another, but AI is uniquely positioned to provoke a murky and irrational discourse. Not only does the technology allow (and to some extent, demand) the advance consideration of countless possible values compromises, but nearly all of them will be what Philip Tetlock calls "taboo trade-offs."[22] Whichever value gets traded off will attract the ire of some constituency and the attention of lawmakers.[23]

We do not expect the abdication/ad-hoc critique problem to be resolved anytime soon. Several years ago, a now well-known ProPublica article revealed the inherent tension between different types of errors and how each impacts members of protected groups.[24] The study had a vast impact on AI scholarship by provoking acknowledgment and debates about how conflicts between different forms of fairness should be resolved, as we will address below.[25] Yet it appears that the broader lesson of the study

---

[21] Philip E. Tetlock, *Coping with Trade-Offs: Psychological Constraints and Political Implications, in* ELEMENTS OF REASON 242 (Arthur Lupia, Mathew D. McCubbins & Samuel L. Popkin eds., 2000) (explaining, with qualifications, that trade-offs are a political liability).

[22] *Id.* at 249-256. Trading off accuracy for equity, or even trading off one form of equity for another, often triggers a moral outrage that makes decision-making repulsive.

[23] AI design for high-stakes decision-making requires "big decisions" that are made with full awareness and where the "choice not made casts a lingering shadow." CASS R. SUNSTEIN, DECISIONS ABOUT DECISIONS: PRACTICAL REASON IN ORDINARY LIFE 36 (2023) (quoting Edna Ullmann-Margalit, *Big Decisions: Opting, Converting, Drifting*, 58 ROYAL INST. PHIL. SUPPLEMENTS 157, 158 (2006)).

[24] Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/86ZY-EWMF].

[25] Andrew Lee Park, *Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing,* UCLA L. REV. (Feb. 19, 2019),

and its surrounding commentary has not been fully acknowledged and integrated into lawmaking. In recent months, as the rising use of ChatGPT and other generative AI tools has made AI policy debates more visible and urgent,[26] there is still little appetite among lawmakers to establish a hierarchy of values.[27]

This Article provides some limited relief. It is a concise and actionable guide for industry self-regulation against the backdrop of the abdication/ad hoc criticism problem. What follows should be treated as guideposts to assist AI developers in creating fair-enough AI, even as the substantive requirements of AI law remain unsettled.[28] The discussion here will be more concrete and utilitarian

---

https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/ [https://perma.cc/NJU7-UPNY]; Matias Barenstein, *ProPublica's COMPAS Data Revisited,* ARXIV 9-12 (July 8, 2019), https://arxiv.org/pdf/1906.04711 [https://perma.cc/2JS3-ELHG] (recalculating recidivism rates using ProPublica's data); Cynthia Rudin, Caroline Wang, Beau Coker, *The Age of Secrecy and Unfairness in Recidivism Prediction,* HARV. DATA SCI. REV., Winter 2020, at 1, 30-32; Julia Angwin & Jeff Larson, *Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say*, PROPUBLICA (Dec 30, 2016, 4:44 PM ET), https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say [https://perma.cc/67ER-H6XJ]; Anne L. Washington, *How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate*, 17 COLO. TECH. L.J. 131, 150-51 (2018); Mayson, *supra* note 5. For discussion of biased treatment and biased objective functions, see *infra* Sections III.B and III.C.

[26] Interest in Artificial Intelligence skyrocketed after OpenAI made ChatGPT available to the public. Google Trends, *AI*, GOOGLE, https://trends.google.com/trends/explore?date=today%205-y&geo=US&q=AI&hl=en [https://perma.cc/J3P6-DLP2] (graphing an increased rate of searches for the term "AI" over the last five years).

[27] *See* references *supra* note 1.

[28] Of course, courts and regulatory agencies may also find this guide helpful as they work through the difficult task of setting legal priorities between fairness goals. Note that there are also several *procedural* elements that are or will likely become legal requirements. These include AI impact assessments, internal or external audits, field testing, and public-use policies. There are several tools and scholarly articles addressing these process requirements. *See* NIST AI RISK MANAGEMENT FRAMEWORK , *supra* note 6, at 21-24. *See generally* Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments,* 35 HARV. J. L. & TECH. 117 (2021) (discussing how algorithmic impact assessments might be implemented as a regulatory strategy); Nicol Turner Lee, Paul Resnick & Genie Barton, *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms,* BROOKINGS (May 22, 2019),

than the frameworks that have been developed by lawmakers, and significantly more pedestrian than the aspirational rhetoric that dominates political discourse. We set out the most popular conceptions of AI fairness and use these as the key ingredients—the sometimes-conflicting objectives—that must be mixed and optimized together in accordance with a priority of values. We argue that unless a statute, regulation, or judicial precedent has explicitly defined how tradeoffs between competing forms of fairness must be resolved, almost any conscientious and well-considered decision for prioritizing values should be considered fair enough. Likewise, as long as an AI application is not needlessly deficient along one of the conceptions of fairness described here, it should be considered fair enough.

We will proceed in a format that could serve as a checklist for an AI ethics officer. In Part I, we encourage the decision maker to ask whether it is appropriate or desirable to differentiate between subjects *at all.* If not, there will be little reason to assume the social costs inherent to differentiation and prediction (lost privacy, a challenge to dignity, and potential for bias). No doubt differentiation will often be necessary, as when an actuary or a criminal-court judge is attempting to assess risk; or when a college or employer is trying to discern talent.

Part II begins the discussion of our taxonomy of unfairness. In this section, we discuss inaccuracy. This is sometimes diminished in, or even distinguished from, discussions of AI fairness, but should not be.

Part III provides a thorough discussion of AI bias, including its different *meanings* and its different *sources*. Our treatment of AI bias makes use of many existing sources, but organizes the concepts

---

https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ [https://perma.cc/SWD8-ZJWS]. We recognize that substantive decisions are improved through good process including use policies and formal risk assessments, and we frequently recommend throughout this article that AI developers document and justify their ethics-related decisions. In this piece, though, we focus on ambiguities and difficulties in deciding on the substance of AI fairness and the range of acceptable options. We (mostly) refrain from comment on the appropriate procedures to reliably make good decisions within that range.

in a way that is more amenable to assessment and correction.[29]

Part IV discusses the perception of disproportionality, when small differences between individuals yield big differences in how they are treated.

Part V considers the potential for gaming: what we refer to elsewhere as gameability—when an algorithm or AI can be exploited through strategic behavior by the individuals or entities that are being judged.

Part VI discusses opaqueness, which interferes with understanding, acceptance, and accountability—and in some instances, undermines basic rights related to autonomy and data protection (at least in the EU).

We conclude with a brief discussion of the "compared to what?" critique—that is, if an AI is determined to be unfair, would outcomes really be *more* fair without it?[30] All of the concerns about AI fairness apply, to greater and lesser degrees, to the "black boxes" of the mind that form our human judgment. AI fairness is not really presenting new problems. It is just unearthing old, festering policy disagreements that had been permitted to sit in a polite, dusty stalemate. When we did not have the means to enumerate and choose between privacy and accuracy, or between bias and explainability, and so forth, decision makers could relax behind the shroud of impossibility and noise. Governance of AI has no such

---

[29] *See, e.g.*, Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC, 2016 O.J. (L 119) 1, 35 [hereinafter GDPR] ("Personal data shall be . . . processed . . . in a transparent matter . . . .").

[30] For this reason, we disagree with the 2023 *AI Now* annual report which, like us, remains skeptical of the value of purely procedural safeguards like self-evaluation and audits, but unlike us, favors bright-line bans. AMBA KAK & SARA MYERS WEST, AINOW, 2023 LANDSCAPE: CONFRONTING TECH POWER 36-37 (2023), https://ainowinstitute.org/wp-content/uploads/2023/04/AI-Now-2023-Landscape-Report-FINAL.pdf [https://perma.cc/4Q5E-6TQ5] (criticizing frameworks that "intervene through process-based modes" that allow for strategic or superficial compliance, and recommending bright line rules instead). The value of bright-line rules would be highly contingent on both a hierarchy of values and on the relative ethical performance of AI compared to the human baseline. Public discourse has not reached a consensus on either of these for most AI applications.

luxury.[31]

Two last notes for completeness's sake: throughout this Article, we assume that AI judgment or decision support is legitimate in principle. This assumption is not without controversy, as many laws or legal proposals have attempted to enshrine the belief that humans have a fundamental right to human decision-making, where appropriate.[32] We will sidestep this debate.[33] Our aim is to help lawmakers and industry steer AI-assisted decision-making to be thoughtful, wherever it occurs, and for debates about the ethics of AI to be intellectually honest.

We also set aside the matter of defining AI. In our discussion below we refer to AI, machine learning, and algorithmic decision-making systems. These terms are not interchangeable, but defining and mapping out the confines of "AI" is a complicated task (almost impossible given the frequent changes in technology) and unnecessary for the purposes of this Article. All automation systems share the attributes—such as lack of human discretion and difficulties in ex-post justification of actions—that are important for this discussion, whether they formally constitute "AI" or not. They all involve an automated process that supplements some form of human discretion at a moment where a decision must be made.

---

[31] Adding to the pressure, AI-governance debates take place against a backdrop of a growing AI industry in China, where AI regulations and guidance are significantly different than in Europe and the United States. Pascale Fung & Hubert Etienne, *Confucius, Cyberpunk and Mr. Science: Comparing AI Ethics Principles Between China and the EU,* 3 AI & ETHICS 505, 510 (2023).

[32] *Blueprint for an AI Bill of Rights*, *supra* note 1, at 46. Scholars have also considered whether there should be a right or requirement to have human involvement or oversight in AI decision-making. I. Glenn Cohen, Boris Babic, Sara Gerke, Qiong Xia, Theodoros Evgeniou & Klaus Wertenbroch, *How AI Can Learn from Law: Putting Humans in the Loop Only on Appeal,* NPJ DIGIT. MED., Aug. 25, 2023, at 1, 1-2. *See generally* Andrea Roth, *Trial by Machine,* 104 GEO. L. J. 1245 (2016) (advocating for "man-machine collaboration" in machine-driven criminal adjudication). For a salient example of this notion in EU law, see GDPR, *supra* note 29, art. 22.

[33] Likewise, we will not address objections to AI systems that are only indirectly related to its recommendations. Specifically, concerns about how input data is sourced or about the financial costs or time needed to run a system are undeniably important factors, but they can and do apply to human systems as well.

I. A Preliminary Question: Is It Fair to Pass Judgment in the First Place?

Decisions about scarce resources and penalties can be made by one of two methods: either pooling potential recipients and distributing the resource using a neutral (or seemingly neutral) factor such as queues, lotteries or taking turns, or by discriminating between potential recipients. In practice, most decisions use a hybrid of these two approaches. The diversity visa lottery, for example, is mostly a pooling system (at least with respect to immigrants from one particular country who meet initial qualifications) because it awards visas by randomly selecting a set number of visa applicants from a particular country.[34] By contrast, a discriminating system would not use random selection, equal apportionment, or queues. Instead, a discriminating factor or process would be premised on the individual's merit, need, skill, or some other measurable (and often predetermined) objective.

Pooling schemes are designed to treat all subjects in the pool the same without assessing the merits or costs associated with any person in the pool. Pooling would be unremarkable for homogenous pools, where everyone is relatively interchangeable. But pooling is also frequently applied to heterogeneous populations. It reflects an implicit policy choice to treat potentially distinguishable people the same and render differentiating factors irrelevant. For example, by prohibiting health insurers from considering preexisting health conditions when defining the terms and price of a health plan, the Affordable Care Act required insurers to ignore factors that would be very relevant to predicted medical costs.[35] By doing so, it converted health insurance from a discrimination scheme to a pooling one. Even though we know ex ante that the pool could be separated into higher-risk and lower-risk subpools, the law compels insurers to ignore this information. In doing so, the law forces the low-risk pool to cross-subsidize the high-risk pool in order to more

---

[34] 8 U.S.C. § 1153(e)(2) (2018); Carrillo-Gonzalez v. INS, 353 F.3d 1077, 1078 (9th Cir. 2003) (discussing the diversity immigration visa lottery); JON ELSTER, LOCAL JUSTICE: HOW INSTITUTIONS ALLOCATE SCARCE GOODS AND NECESSARY BURDENS 57–59, 72 (1992) (same).

[35] Digital Communications Division, *Can I Get Coverage if I Have a Pre-Existing Condition?*, U.S. DEP'T OF HEALTH & HUM. SERVS. (Apr. 20, 2023), https://www.hhs.gov/answers/health-insurance-reform/can-i-get-coverage-if-i-have-a-pre-existing-condition/index.html [https://perma.cc/VRE3-A7VL].

broadly spread the costs of care for patients who are ill (or are predisposed to become ill) and thus broadly share risk throughout society (at the cost of allowing some moral hazard to unfold).

In contrast to pooling, discrimination schemes do not even attempt to treat all subjects the same. Because the term "discrimination" is often used synonymously with unlawful discrimination based on race, sex, or other characteristics, we will use the term "differentiation" for clarity's sake. Differentiation schemes are meant to allocate resources based on differences in individuals' actual traits *or* predicted behavior. The goal for differentiation is to allocate a resource based on an abstract and fundamentally unknowable quality of the subjects relating to their skill, risk, need, merit, or some other quality that is appropriate to the relevant setting and goals. We will call this quality the "ultimate goal."[36] For college admissions officers, the ultimate goal might be a mix of raw intelligence and future career success, thus promoting the school's reputation. For creditors, the ultimate goal is the subject's ability to pay their debt in the future. Oftentimes, a secondary goal of a differentiation scheme is to create incentives for the subjects of the scheme to meet their personal potential with respect to this important quality.

Predictions of the future are fundamentally unknowable—even the most concrete and value-neutral ultimate goals can only be estimated. For instance, consider the notion of "impairment"—an element in crimes related to driving while intoxicated. The abstract notion of impairment is proved through concrete measures of intoxication, which usually approximate the proportion of alcohol in the subject's breath or blood using the most sensitive and accurate equipment available. This is a very close substitute for "impairment," but not perfect. And it features systemic and predictable differences between individuals which might render it colloquially "unfair." Yet it is broadly accepted.

When a decision-maker is trying to differentiate between

---

[36] There is no universal term for this idea—the hard-to-observe goals that usually require a proxy for the purpose of measurable assessment. We will use "ultimate goal" because it is a term that is often used in discussions of clinical trials. *E.g.*, Wendy J. Coster, *Making the Best Match: Selecting Outcome Measures for Clinical Trials and Outcome Studies*, 67 AM. J. OCCUPATIONAL THERAPY 162, 163 (2013); Michelle Nottage & Lillian L. Siu, *Principles of Clinical Trial Design*, 20 J. OF CLINICAL ONCOLOGY 42s (Sept. Supp. 2002).

subjects to allocate resources based on an ultimate goal—deciding, for example, that "impaired" drivers can be punished but unimpaired ones must be released, or that the most "education-maximizing and college-ready" applicants should be offered admission to the university—a decisionmaker *must* use an algorithm. That is, they must adhere to some set of rules to weight and balance the proxies that the decisionmaker believes are well correlated with the ultimate goal. This is true whether the decision-maker is relying partially or entirely on computers or is conducting the analysis entirely through humans.[37] Human differentiators may be less consistent than a machine would be by diverging from their intended procedures, by adding noise, or by changing the ultimate goal that they are striving to achieve, but they cannot escape the fact that proxies are used to evaluate and ultimately decide upon something unobservable.

We raise the topic of pooling and differentiating because a preliminary analysis that often gets lost in the discussion of AI is whether a resource really should be distributed on a differentiating, rather than pooling, basis. To be sure, the answer will often be "yes, we need to differentiate." Resource allocation often requires differentiating between subjects in a system that assesses merit, need, or some other key factor to drive the allocation of the resource. It may be a human rather than a machine doing the differentiating, but differentiating there will be.

AI policymakers (and those critiquing their work) should not skip this step. This preliminary question about whether the resource can be allocated randomly, shared or divvied relatively equally helps to disqualify decision-making systems that do not benefit enough from a discriminating algorithm at all, whether implemented by

---

[37] However, a human system of algorithmic decision-making is more likely to use inconsistent objective functions that change depending on time and the person making the decision. *See generally* Todd McElroy, Joanna Salapska-Gelleri, Kelly Schuller & Martin Bourgeois, *Thinking About Decisions: How Human Variability Influences Decision-Making, in* BRAIN, DECISION MAKING AND MENTAL HEALTH 487 (Nima Rezaei ed. 2023) (discussing several factors that influence human decision-making). This is problematic if one of the goals of the decision-making is consistency, but if humans are using multiple but equally valid objectives, a human system may be more pluralistic (at the cost of consistency.) *See* Edward Parson, Robert Lempert, Ben Armstrong, Evan Crothers, Chad DeChant & Nick Novelli, *Could AI Drive Transformative Social Progress? What Would This Require?*, AI PULSE (Sept. 26, 2019) (discussing human and AI decision-making that incorporates multiple criteria and objectives).

human or machine. A responsible data steward might prefer to use a lottery or a queueing system if there is little separating the "deserving" from the rest. Or she might even decide to do so when the true regulatory preference is to apply a merit- or desert-based system, yet the outcome measures that she would use to approximate "merit," "desert," or some other unobservable objective, are so noisy that a system of separation is unsound. In other instances, pooling mechanisms might be chosen to honor equality over whatever benefits might come from differential treatment.

II. Unfairness as Inaccuracy

Accuracy problems can cause an algorithm to make decisions in ways that are not only inefficient but patently unfair.[38] It is morally wrong for algorithms used by state agencies or private actors to make important decisions affecting legal rights and privileges on the basis of flawed or error-prone assessments.[39] American law captures this to some extent by requiring agencies to avoid decision-making that is "arbitrary and capricious."[40] A sudden change in how various factors are weighted and considered creates a random, nonrational system that is antithetical to any plausible policy.[41] Likewise, an

---

[38] Assessing accuracy requires a reliable means of determining, eventually, whether the output of an AI system was right or wrong. The International Organization for Standardization defines accuracy as the "closeness of results of observations, computations, or estimates to the true values or the values accepted as being true." *Trustworthiness Vocabulary*, INT. ORG. STANDARDIZATION (2022), https://www.iso.org/obp/ui/#iso:std:iso-iec:ts:5723:ed-1:v1:en [https://perma.cc/M886-NJAX].

[39] DANIEL KAHNEMAN, OLIVIER SIBONY, CASS R. SUNSTEIN, NOISE: A FLAW IN HUMAN JUDGMENT 6 (2021) (describing the amount of noise in most important decision-making systems as "scandalously high").

[40] 5 U.S.C. § 706(2)(A) (2018); Aziz Z. Huq, *Constitutional Rights in the Machine Learning State*, 105 CORNELL L. REV. 1875, 1908-09 (2020) (discussing Due Process Clause constraints on arbitrary decision-making by government). We discuss the application of the "arbitrary or capricious" standard's application to AI at length in Jane R. Bambauer, Tal Zarsky & Jonathan Mayer, *When a Small Change Makes a Big Difference: Algorithmic Fairness Among Similar Individuals*, 55 U.C. DAVIS L. REV. 2337, 2376-77 (2022). Note, however, that the *Chevron* doctrine was recently overturned in *Loper Bright Enterprises v. Raimondo*, 144 S.Ct. 2244, 2273 (2024).

[41] *Cf.* Dep't of Homeland Sec. v. Regents of Univ. of Cal., 591 U.S. 1, 22 (2020)

inaccurate algorithm is unfair given its breach of several equality principles—chief among them the principle of "equal treatment of equals." Those with similar attributes and factors relevant to the allocation goals should receive similar treatment, and similar chances of errors.

The law does not usually impose a minimum threshold of accuracy for hiring, credit, or other market decisions, even when they are made on random or sentimental bases. In employment, for example, the standard approach in at-will employment states is that hiring and retention decisions can be made on any basis so long as they are not made on the basis of a discrete set of prohibited factors such as race or an accommodatable disability.[42] The theory behind this doctrine is not that employment is wholly divorced from merit but that the market will do enough to discipline employers without legal intervention. Nevertheless, even if there is no legal mandate to value accuracy in a scoring or decision-making algorithm, broadly shared notions of fairness and desert will be violated if a decision-making system is too flawed.[43] This might lead to bad press, regulatory backlash or even internal unrest—all unwanted outcomes for a profit-seeking firm.

Note that some scholars in AI fairness treat accuracy or efficiency as distinct from fairness. In their book *The Ethical Algorithm,* for example, Michael Kearns and Aaron Roth define fairness to include respect for privacy, avoidance of bias, gaming, and overfitting, and explainability, and contrasts each of these with the goal of basic performance.[44] This sets up an "accuracy versus fairness" rhetorical battle that we think is misleading. By recognizing the importance of accuracy for a just and fair system of resource allocations, we properly see the "accuracy versus fairness" debate as one specific cross-section of a larger "fairness versus fairness versus fairness" debate.

(explaining that a policy "[cannot] be rescinded in full 'without any consider whatsoever'" of an alternative policy (quoting *Motor Vehicle Manufacturers Association of the United States, Inc. v. State Farm Mutual Automobile Insurance Co.*, 463 U.S. 29, 51 (1983)).

[42] 42 U.S.C. § 12112(a) (2018); 42 U.S.C. § 2000e-2(a) (2018).

[43] NIST AI RISK MANAGEMENT FRAMEWORK , *supra* note 6, at 13-14 (discussing validity, reliability, and accuracy as the foundation in trustworthy AI).

[44] *See* KEARNS & ROTH, *supra* note 2, at 74-78, 208.

III. Unfairness as Bias

Perhaps the greatest concern animating exploration and discussion of algorithm fairness revolves around unintended biases that can be quietly imbedded into AI.[45] When the biases appear in automated algorithms, they can exacerbate existing racial, gender, or class disparities (or, at least, fail to improve them). This often occurs despite the effort and good intentions of programmers.

Most AI frameworks require the developers or intended users of an algorithm to proactively assess whether it will cause discriminatory or inequitable results for a historically disadvantaged group.[46] An assessment will begin by looking for disparities across race, gender, and other identity categories. But this is just the start. AI ethics officers should then figure out why the disparities emerge.

---

[45] *See* Zeynep Tufekci, The Real Bias Built In at Facebook, N.Y. TIMES (May 19, 2016), https://www.nytimes.com/2016/05/19/opinion/the-real-bias-built-in-at-facebook.html [https://perma.cc/6WD5-VMMZ] (discussing how programmers often cannot predict the outcomes of their programs); Latanya Sweeney. *Discrimination in Online Ad Delivery*, COMMC'NS OF THE ACM (May 1, 2013), at 44-54; Hannah Devlin, *Discrimination by Algorithm: Scientists Devise Test to Detect AI Bias*, THE GUARDIAN (Dec. 19, 2016, 2:30 AM EST), https://www.theguardian.com/technology/2016/dec/19/discrimination-by-algorithm-scientists-devise-test-to-detect-ai-bias [https://perma.cc/6AYR-A4CS].

[46] Exec. Order No. 14,091, 88 CFR §§ 10825, 10831-32 (2023) (defining "equity" as: "[T]he consistent and systematic treatment of all individuals in a fair, just, and impartial manner, including individuals who belong to communities that often have been denied such treatment, such as Black, Latino, Indigenous and Native American, Asian American, Native Hawaiian, and Pacific Islander persons and other persons of color; members of religious minorities; women and girls; LGBTQI+ persons; persons with disabilities; persons who live in rural areas; persons who live in United States Territories; persons otherwise adversely affected by persistent poverty or inequality; and individuals who belong to multiple such communities."); *Blueprint for an AI Bill of Rights*, *supra* note 1, at 26-27. The Blueprint recommends conducting assessments without specifying what type of bias should be assessed, or what minimum thresholds should be used. *Id.* at 27 ("Automated systems should be tested using a broad set of measures to assess whether the system components [] produce disparities . . . . Disparities that have the potential to lead to algorithmic discrimination, cause meaningful harm, or violate equity goals should be mitigated."). The Blueprint adopts the same list of disadvantaged groups used in Executive Order No. 14,091. *Id.* at 26.

There are four distinct (and oftentimes mutually exclusive[47]) categories of bias: biased output error, biased treatment error, biased objective functions, and disparate impact. Our organization of bias differs from the categories of bias often described in the technical literature,[48] but offers more precision and better organization for the purposes of policy considerations. We explain each category of bias below.

A. Biased Output Error

The most straightforward form of AI bias comes in the form of output error. All predictive systems will have error—some difference between predicted results and actual results. That error is a biased one if it is larger or directionally skewed to the detriment of a protected group.[49] Well before the literature on AI bias began to take shape, experts often detected bias in human systems by looking for differences in the success or failure rates for members of different groups who are scored or treated the same way.[50] This is still the first and most generally agreed-upon form of evidence that a system is biased or discriminatory. For example, in *Floyd v. NYPD*, the case challenging the practices of the NYPD stop and frisk program, the fact that frisks of African Americans were less likely to produce a weapon than frisks of whites showed that police were using a different standard for minorities that resulted in stopping them with less cause than their white counterparts.[51]

---

[47] Mayson, *supra* note 5, at 2223 ("[T]here are many possible metrics of racial equity in statistical prediction, and some of them are mutually exclusive."); *see* Kleinberg et al., *supra* note 16, at 4-6 (setting out three forms of bias and characterizing the inherent tradeoffs between different forms of what we are calling "biased output error.").

[48] NIST, for example, divides bias across three overlapping categories: systemic bias, computational/statistical bias, and human-cognitive bias. NIST AI RISK MANAGEMENT FRAMEWORK, *supra* note 6, at 18; REVA SCHWARTZ, APOSTOL VASSILEV, KRISTEN GREENE, LORI PERINE, ANDREW BURT & PATRICK HALL, NAT'L INST. OF STANDARDS & TECH., TOWARDS A STANDARD FOR IDENTIFYING AND MANAGEING BIAS IN ARTIFICIAL INTELLIGENCE 6-11 (2022), https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf [https://perma.cc/6A5R-TUGQ].

[49] For a discussion of different types of error measurements, each of which could contain disparities for some groups, see Mayson, *supra* note 5, at 2243-46.

[50] *See* GARY S. BECKER, THE ECONOMICS OF DISCRIMINATION 101-134 (2d ed. 1971) (describing statistical analyses where treatment effects were different for different racial groups).

[51] Floyd v. City of New York, 959 F. Supp. 2d 540, 559, 562 (S.D.N.Y. 2013).

Suppose, for example, that a scoring algorithm was designed to predict performance on a math test based on a variety of school attendance, grades, and test performance data.[52] When the students take the math exam, their actual scores can be compared to their predicted scores. If the predicted scores for female students were, on average, 2 points lower than their actual scores, and if the predicted scores for male students were, on average, 0.3 points higher than their actual scores, the algorithm will have exhibited biased errors for women. Any use of the predictions to direct resources or rewards could have been biased as a result.[53]

The main sources of bias in output errors are (1) poor quality training data; and (2) constrained models.

1. Insufficient or Unrepresentative Training Data

Many examples of algorithm bias can be explained by flawed data during the machine-learning training process.[54] Algorithms that are derived using a small training dataset will have accuracy problems across the board, and those problems are typically worse for small subpopulations: even if the training set is large, biased error can occur when a group is underrepresented in the training data as compared to their proportion in the relevant population. This was

---

[52] It may seem like a silly or contrived example to imagine an algorithm that would predict how a student would do on an exam that they have not yet taken, but the United Kingdom chose to do precisely this in 2020, when COVID interfered with students' A-level exams. Daan Kolkman, "*F**k the Algorithm"?: What the World Can Learn from the UK's A-Level Grading Fiasco,* LONDON SCH. OF ECON. BLOG (Aug. 26, 2020), https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/ [https://perma.cc/39BG-49H2].

[53] This example involves both outputs and true results that are continuous variables (math scores on a 0-100 scale). Errors for outputs that attempt to predict binary or discrete true results must be measured as false-positive rates (also known as specificity error) or false-negative rates (sensitivity error). Mariska M.G. Leeflang, Karel G.M. Moons, Johannes B. Reitsma & Aielko H. Zwinderman, *Bias in Sensitivity and Specificity Caused by Data-Driven Selection of Optimal Cutoff Values: Mechanisms, Magnitude, and Solutions,* 54 CLINICAL CHEMISTRY 729, 729 (2008).

[54] Nicol Turner Lee, *Detecting Racial Bias in Algorithms and Machine Learning*, 16 J. INFO. COMM. & ETHICS IN SOC. 252, 256 (2018); *Shedding Light on AI Bias with Real World Examples*, IBM (Oct. 16, 2023), https://www.ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples [https://perma.cc/YV2U-KEW8].

a problem, at least for a time, with facial recognition algorithms trained on datasets where images of black faces were underrepresented.[55] But overrepresentation within the training data can *also* cause disparate error rates by causing overfitting problems: for instance, when a minority is oversampled in stops and frisks, that population will be overrepresented in the number of crimes detected.[56] Thus, the best practice for AI development is to use a robust and representative dataset for training that would not require any significant oversampling.

2. Constrained Models

Sometimes, a prediction algorithm will feature greater errors for some demographic groups for reasons that are mysterious or that cannot be accounted for with the available data. Assessments could reveal that the outputs for one subgroup are predictably and reliably wrong. This bias can be mitigated by allowing the AI model to take demographic information into account. As a result, AI ethicists and legal academics have come around to recommending that machine-learning systems have access to race, gender, and other sensitive attributes if their inclusion will provide a helpful correction.[57] For example, as Nicol Turner Lee explained,

> If an algorithm is forbidden from reporting a different risk assessment score for two criminal defendants who differ only in their gender, judges may be less likely to release female defendants than male defendants with equal actual risks of committing another crime before trial. Thus, blinding the algorithm from any type of sensitive

---

[55] *See* Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge & Anil K. Jain, *Face Recognition Performance: Role of Demographic Information*, 7 IEEE TRANSACTIONS ON INFO. FORENSICS & SEC. 1789, 1791, 1798, 1800 (2012). However, facial recognition software has become much more accurate and less biased in the intervening decade. *See generally* Patrick Grother, Mei Ngan, Kayee Hanaoka, Joyce C. Yang & Austin Horn, *Face Recognition Technology Evaluation (FRTE) Part 1: Verification*, NAT. INST. OF STANDARDS & TECH. (Jan. 30, 2025), https://pages.nist.gov/frvt/reports/11/frvt_11_report.pdf [https://perma.cc/7MJB-WA4Z] (summarizing the most performance data for facial-recognition algorithms).

[56] *See* BERNARD HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE 30-31 (2007).

[57] Mayson, *supra* note 5, at 2263-67.

attribute may not solve bias.[58]

However, a company that intentionally includes race, gender, or any other protected characteristic, even for the reasons of enhancing equity, will run a significant legal risk under current discrimination laws and norms.[59] Indeed, far from including race and gender variables, AI and algorithm developers often go out of their way to not only strip away race from structured training and input data, but even to avoid variables that are highly correlated with race.[60]

3. Unavoidable Biased Error

AI developers should make reasonable efforts to avoid these common sources of bias, and regularly test for differences across various demographic groups in false positive rates, false negative rates, or some weighted combination of the two.[61] The reasons that low-quality training data may be selected, or poor modeling may occur, can itself have myriad causes including inadequate education and supervision, haphazard engineering protocols, and a lack of diversity among the staff.[62] All of these potential pitfalls should be managed as well as possible.

However, even when the team is well composed and well trained, there are some practical limitations on measuring and

---

[58] Lee et al., *supra* note 28.

[59] *See Blueprint for an AI Bill of Rights*, *supra* note 1, at 26 ("Directly using demographic information in the design, development, or deployment of an automated system (for purposes other than evaluating a system for discrimination or using a system to counter discrimination) runs a high risk of leading to algorithmic discrimination and should be avoided."); 42 C.F.R. § 92.210 (2024) (prohibiting covered entities from "discriminat[ing] on the basis of race, color, national origin, sex, age, or disability in its health programs or activities through the use of patient care decision support tools," which include AI algorithms); Todd Feathers, *Major Universities Are Using Race as a "High Impact Predictor" of Student Success*, THE MARKUP (Mar. 2, 2021, 8:00 AM ET), https://themarkup.org/machine-learning/2021/03/02/major-universities-are-using-race-as-a-high-impact-predictor-of-student-success [https://perma.cc/AXM5-BPVJ] (describing the use of race in predictive systems to target students at a higher risk of drop-out with more resources).

[60] *Blueprint for an AI Bill of Rights*, *supra* note 1, at 26; Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 677-93 (2016).

[61] *See* Mayson, *supra* note 5, at 2243-2248 (examining different measures for evaluating false-negative, false-positive, true-negative, and true-positive rates).

[62] Michael Li, *To Build Less-Biased AI, Hire a More-Diverse Team,* HARV. BUS. REV. (Oct. 26, 2020), https://hbr.org/2020/10/to-build-less-biased-ai-hire-a-more-diverse-team [https://perma.cc/6Q2B-NU5L].

closing group differences in error rates. The observation of error rates requires some adequate means of observing what the "true" outcome is, or more accurately, what it *would have been* if they had not been affected by the algorithmic treatment. These counterfactuals can be difficult not only because of limits on observing scored individuals who are not affected by treatment, but also because the true "outcome" might not be observable by *anybody*. For example, it is easy to discuss the contrived example with predicted and actual math test scores broken down by gender,[63] but much harder to measure biased error in the real-life context of an issue like loan grant rates, where we must ask what would have happened if a loan applicant who was denied a loan had actually received it.

Next, even if assessment is feasible, complete equality of error can be difficult, if not mathematically impossible to achieve: in any heterogenous community, an AI cannot have equal error rates across all groups and all types of error.[64] Thus, AI ethics officers will have to make thoughtful and well-documented decisions about which type of output error is the most problematic, and which populations are the most vulnerable in the context in which the algorithm will be used. But the impossibility of equal error is no reason to reject AI systems altogether; after all, human decision makers will also have bias and error.

B. Biased Treatment Without Biased Output Error

Even if the output of an algorithm is bias free (as in, about equally accurate for each value across demographic categories), the system as a whole can nevertheless produce bias in how it treats members of different groups. A human user of an algorithmic decision-making system can turn neutral output into inequitable treatment by (1) interpreting the results or overriding them in a way that introduces error and bias; or (2) assigning consequences to each output that create harsh cutoffs in a place in the distribution that has particularly strong effects for minority or disadvantaged groups.

1. Biased Humans in the Loop

AI systems are often executed with a human-in-the-loop who can override a strange or clearly wrong recommendation. These

---

[63] *See* notes 44-45 and accompanying text.
[64] Kleinberg et al., *supra* note 16, at 8.

"cyborg" systems can sometimes outperform AI alone, but there is no guarantee.[65] A human user of an algorithmic decision-making system can turn neutral output into inequitable treatment by interjecting their own flawed judgment. For example, a study of risk scores used by child-protective-service centers found that when human decision makers deviated from the recommendation of a scoring system, they tended to screen more Black families into the high-risk treatment.[66] And a landmark study by Megan Stevenson and Jennifer Doleac found that judges tended to deviate from risk score recommendations in a manner that led to shorter sentences for younger defendants and harsher treatment for Black defendants.[67]

Counterintuitively, human judgment may also be harsher across the board, for all demographics, as several studies now suggest that criminal sentences and judicial determinations become somewhat less harsh (more lenient) when risk scores are introduced.[68] The

---

[65] Catherine Pope, Susan Halford, Joanne Turnbull & Jane Prichard, *Cyborg Practices: Call-Handlers and Computerised Decision Support Systems in Urgent and Emergency Care*, 20 HEALTH INFORMATICS J. 118, 123-24 (2014); *cf.* Donna J. Haraway, *A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century*, *in* SIMIANS, CYBORGS, AND WOMEN: THE REINVENTION OF NATURE 149 (Donna J. Haraway ed., 1991) (conceiving of a "leaky distinction" between "animal-human . . . and machine").

[66] Alexandra Chouldechova, Emily Putnam-Hornstein, Diana Benavides-Prado, Oleksandr Fialko & Rhema Vaithianathan, *A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions*, 81 PROC. MACH. LEARNING RSCH. 1, 7 (2018) (citing Alan J. Dettlaff, Stephanie L. Rivaux, Donald J. Baumann, John D. Fluke, Joan R. Rycraft & Joyce James, *Disentangling Substantiation: The Influence of Race, Income, and Risk on the Substantiation Decision in Child Welfare,* 33 CHILD. & YOUTH SERVS. REV. 1630, 1632-1634 (2011)).

[67] Megan Stevenson & Jennifer Doleac, *Algorithmic Risk Assessment in the Hands of Humans*, 16 AM. ECON. J.: ECON. POL'Y 382, 383-84 (2024) ("We run a series of tests to see which demographic factors predict deviating upwards or downwards in sentencing. We find that, conditional on the risk score, judges are substantially more lenient with young defendants than older defendants and substantially harsher on Black defendants than non-Black defendants."); *see also* Megan T. Stevenson & Christopher Slobogin, *Algorithmic Risk Assessments and the Double-Edged Sword of Youth,* 96 WASH. U. L. REV. 681, 683-84 (2018) (describing the tension between youth as a risk factor and as a lenience factor in criminal justice).

[68] Stevenson & Doleac, *supra* note 67, at 383 ("Sentencing by algorithm would have resulted in a sharp decrease in both the probability of incarceration and the

corollary is that judges are harsher when they are permitted or required to make their own independent assessments, or are easily able to override the algorithmic recommendation. These findings demonstrate the importance of effective training for end users or "humans-in-the-loop" within an AI system who have an opportunity to influence how the subjects of the prediction tool are treated. The studies also challenge the wisdom of opt-out and human appeals procedures that have been incorporated into many AI regulatory proposals such as the White House *Blueprint for an AI Bill of Rights*[69] or the California Privacy Protection Agency's proposed rules for AI.[70] These popular and well-intentioned release valves may provide some accountability and relief for the misjudged, but they also could come at a significant cost to accuracy and equality.

2. Insufficient Graduation in Treatment

Even when humans do not supplant machine judgment with their own, bias can still be introduced when treatment or consequences are assigned to AI outputs. Consider what is perhaps the most widely discussed example of algorithmic bias—the ProPublica study of COMPAS recidivism risk scores which we briefly noted above.[71] These scores are used to determine whether a criminal defendant should be detained before trial—involved a scoring system that did not actually exhibit bias in its outputs.[72]

The study involved a cohort of individuals who were arrested

---

sentence length. In practice, however, judges diverted far fewer individuals [to noncarceral programs] than were recommended by the algorithm."); *see also* Megan T. Stevenson & Jennifer L. Doleac, *The Counterintuitive Consequences of Sex Offender Risk Assessments at Sentencing,* 73 U. TORONTO L. J. SUPP. 59, 68-69 (2023) (observing decreased sentences for those convicted of rape when Virginia judges used a risk assessment tool); Megan Stevenson, *Assessing Risk Assessment in Action,* 103 MINN. L. REV. 303, 327 (2018) (concluding that "it remains reasonable to think that a well-built actuarial tool can out-predict a judge on future offending" but that "the margin of improvement remains unclear"). For a more AI-skeptical (or pro-human) perspective, see Daniel Solove & Hideyuki Matsumi, *AI, Algorithms, and Awful Humans,* 96 FORDHAM L. REV. 1923, 1925 (2024).

[69] *Blueprint for an AI Bill of Rights*, *supra* note 1, at 46.

[70] CAL. PRIV. PROT. AGENCY, *supra* note 6.

[71] Angwin et al., *supra* note 24.

[72] *Id.* See Prathamesh Patalay, *COMPAS: Unfair Algorithm?*, MEDIUM (Nov. 21, 2023), https://medium.com/@lamdaa/compas-unfair-algorithm-812702ed6a6a [https://perma.cc/LF5V-A8XW] for a summary of criticism.

and charged with crimes and, importantly, scored using the COMPAS recidivism algorithm. However, they were all released for exogenous reasons. This allowed the researchers to compare the COMPAS scores predicting the chance of reoffending with the actual rearrest data.

The scores themselves were not biased in the manner described in the last section—there was parity across scores, insofar as a Black defendant who received a particular score had the same likelihood of rearrest as a white defendant.[73] However, the study found that the scores were biased by looking at the data from another direction. Among those who did not reoffend, Black nonoffenders were assigned higher risk scores than white nonoffenders.[74] If judges had decided to detain defendants assigned to middle-to-high scores, Black nonoffenders would have been detained at almost twice the rate as white nonoffenders.[75] Conversely, white defendants who *did* reoffend had lower scores, and would have been released more often than Black reoffenders (again by a factor of two). This surprising result, where nonbiased outputs can nevertheless yield biased treatment, is a matter of distribution[76]:

---

[73] William Dieterich, Christina Mendoza & Tim Brennan, *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity,* NORTHPOINTE INC. 10-13 (July 8, 2016), https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf [https://perma.cc/8V7U-LJ5U].

[74] Angwin et al., *supra* note 24.

[75] *Id.*

[76] Kenny Kyunghoon Lee, *Replicating Propublica's COMPAS Data Analysis with Python,* NUMERICAL THOUGHTS (Nov. 1, 2020), https://blog.kennylee.info/projects/python/data/machinelearning/bias/2020/11/01/analyze-Compas.html [https://perma.cc/LQT6-5GKP].

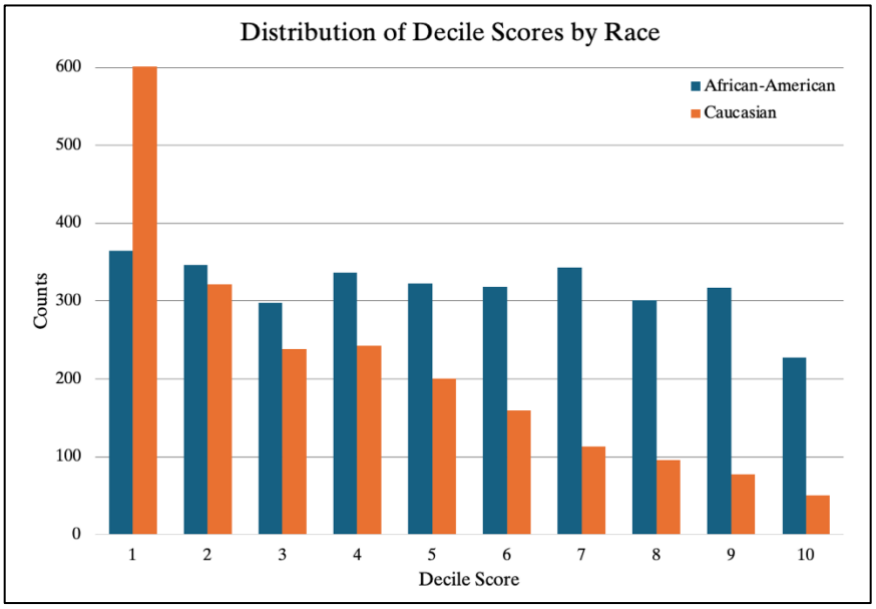

Figure 1: Distribution of COMPAS decile scores by race.-[77]

While white defendants disproportionately fell into the lowest-risk decile, Black defendants had scores that were much more evenly distributed through the middle and high portions of the range, and these are the scores for which the algorithm is least accurate—in other words, for which rearrest predictions are a crapshoot.

Consider the public results of a study of New York's use of COMPAS scores, which used a two-year look-back period to determine whether defendants reoffended.[78] These results come from a different jurisdiction, but are consistent with the data used by ProPublica and easier to understand, based on how it has been graphed.

---

[77] *Id.*

[78] SHARON LANSING, N.Y. STATE DIV. OF CRIM. J. SERVS., NEW YORK STATE COMPAS-PROBATION RISK AND NEED ASSESSMENT STUDY: EXAMINING THE RECIDIVISM SCALE'S EFFECTIVENESS AND PREDICTIVE ACCURACY 3 (2012), https://www.criminaljustice.ny.gov/crimnet/ojsa/opca/compas_probation_report_2012.pdf [https://perma.cc/4CSZ-J8A9].

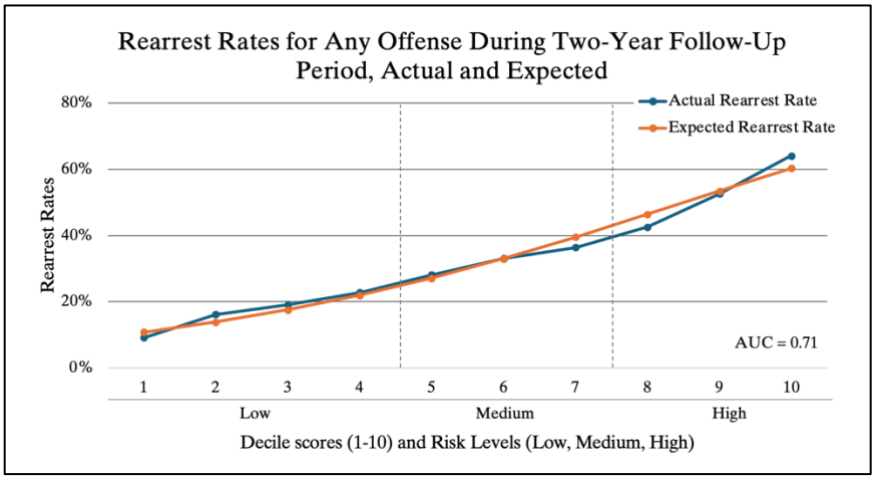

Figure 2: Rearrest rates for any offense during two-year follow-up period: actual and expected rates.[79]

Defendants assigned to scores between 5 and 10 reoffended within two years at rates that ranged from 28.1% to 64.1%.[80] Put another way, 35.9% to 71.9% of the defendants rated medium-to-high risk did *not* reoffend. The defendants scoring between 5 and 7 are particularly difficult to assign consequences to, since in this range, there is a nontrivial risk of reoffending, but most defendants will not do so. Since a much greater portion of Black defendants from the ProPublica study received scores in that 5-to-7 range than did their white counterparts, they would have been disproportionately affected by false-positive treatments. That is, if the judges had chosen to detain defendants with those scores (as the ProPublica study presumed they would), the foreseeable error would have predictably fallen disproportionately on Black defendants.

But nothing about this treatment is inevitable. Judicial systems can and often do use graduated precautions that are better matched to the risk posed. Say a judge released without bail defendants with a score in the 1-to-4 range, released on bail defendants in the 5-to-6 range, released with electronic monitoring defendants in the 7-to-8 range, and detained only defendants with scores of 9 or 10. A "what if" counterfactual study of the sort ProPublica conducted would have starkly different results. Black nonoffenders could still be overrepresented in the group detained (the 9 and 10 scores), most of

[79] *Id.*
[80] *Id.* at 7 fig.1.

the false positives that ProPublica found (coming from the 4-to-8 scores) would evaporate.). Many non-offenders *and* reoffenders, of all races, would have been treated with bail or monitoring instead, and the analysis would have had to be much more nuanced and complex. Bail and monitoring impose some costs, but far fewer than corporal detention.

The larger point is that the weight and meaning assigned to each type of output must incorporate value judgments about what types of error are going to have demographic disparities. ProPublica's notion of fairness is at odds with the notion of fairness that Northpointe used in developing its algorithm.[81] Treatment errors of some sort will be unavoidable, but which are worse, and for whom?[82] For example, in the context of recidivism-risk scores, it is natural to assess racial bias from the perspective of the arrested individuals being scored. For this group, false-positive error is more damaging to the arrestee than false-negative error. But from the perspective of potential crime victims, the reverse is true. Falsely concluding that a defendant arrested for domestic violence is not at risk of reoffending presents more physical risk to the DV victim than falsely concluding that a defendant *is* a risk.[83] When viewed through the lens of potential future victims, the ProPublica findings would suggest that COMPAS offers more protection against wrongful release to Black people than to white people. This is because false-negative error disproportionately leads to the release of violent white arrestees, who then would typically commit crimes against

---

[81] *See* Sam Corbett-Davies, Emma Pierson, Avi Feller & Sharad Goel, *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear.*, WASH. POST (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ [https://perma.cc/5YZA-H4H7] (discussing the differing notions of fairness).

[82] The point we are making here differs from discussions of deviating from a single-threshold rule. That rule requires the same threshold or cut score to be used for every demographic group. Mayson, *supra* note 5 at 2240; Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice,* 68 DUKE L. J. 1043, 1116 (2019). Instead, we argue that it may be appropriate to *decide* where to place the thresholds that will be used (for everybody) based on how small changes in the threshold will affect members of different groups.

[83] *See* Richard Berk, *Accuracy and Fairness for Juvenile Justice Risk Assessments,* 16 J. EMPIRICAL LEGAL STUD. 175, 180-84 (2019).

white victims.[84]

Thus, selecting which type of error matters most, and to which groups of affected parties, is a critical prerequisite to assigning cut scores, treatments, and consequences that will mitigate bias.[85]

Acknowledging that unfairness might result from harsh cutoffs in treatment comes with several concrete implications. At times, this issue might be avoidable when setting systems in place. Are cutoffs a practical necessity? Or can a more graduated treatment be used? Mitigations can come in the design stage. Alternatively, if cutoffs are necessary, choosing the right cutoff criteria could benefit from educating and instructing users about the implications of choosing one criterion or another, and the benefits of using more granular differences in treatment where possible.

As was the case with biased error, bias in treatment is just as likely to occur from purely human systems as computer-assisted ones. In fact, machine systems could do a better job modulating cutoff points where they will have the lowest impact on race or gender disparities in outcomes.

---

[84] Most violent crime occurs intrarace, and domestic violence is no exception. Although interracial couples are somewhat more likely to have incidents of domestic violence, *see* Rachel A. Fusco, *Intimate Partner Violence in Interracial Couples: A Comparison to White and Ethnic Minority Monoracial Couples*, 25 J. INTERPERSONAL VIOLENCE 1785, 1793 (2010), more than eighty percent of opposite-sex couples are monoracial. Gretchen Livingston & Anna Brown, *Intermarriage in the U.S. 50 Years After Loving v. Virginia*, PEW RSCH. CTR. (May 18, 2017), https://www.pewresearch.org/social-trends/2017/05/18/intermarriage-in-the-u-s-50-years-after-loving-v-virginia/ [https://perma.cc/KZA5-9FSY]. Some might think this is an important frame for thinking about the costs of error in pretrial detention, especially since Black women already suffer from a disproportionate share of intimate-partner violence. N.Y. CITY MAYOR'S OFF. TO END DOMESTIC & GENDER-BASED VIOLENCE, 2020 REPORT ON THE INTERSECTION OF DOMESTIC VIOLENCE, RACE/ETHNICITY AND SEX 6 (2021), https://www.nyc.gov/assets/ocdv/downloads/pdf/endgbv-intersection-report.pdf [https://perma.cc/U4FX-XJ4H] (finding that Black female residents of New York were almost four times as likely as other female residents to be a victim of intimate-partner felony rape).

[85] The importance of assigning the most appropriate consequences, and the difficulty of managing the harsh cutoffs, is a well-worn theme in law as well. LEE ANNE FENNELL, SLIDES AND LUMPS: DIVISION AND AGGREGATION IN LAW AND LIFE 7 (2019); Adam J. Kolber, *Smooth and Bumpy Laws,* 102 CALIF. L. REV. 660, 676 (2014).

3. Unavoidable Bias in Treatment

Just as AI outputs cannot avoid every possible type of bias for every sort of error and for every group or combination of attributes,[86] protocols for how to handle or treat different scored individuals or situations will also unavoidably contain some bias. As with every ethical dimension we discuss, responsible AI developers should assess as best as possible, choose their preferred option, and document and justify their design choices.

This will not save an AI developer from criticism or regulatory investigation. For example, a review of AI bias published in the *New England Journal of Medicine* criticized one set of cardiac-risk-scoring systems for disproportionately assigning Black patients to lower-risk treatment options and in the next paragraph, criticized another set of cardiac-risk-scoring systems for assigning Black patients disproportionately to higher-risk options.[87] Each of the paragraphs, standing on its own, makes a good point. Sometimes it is better for a patient to be treated as less healthy (so that they access a treatment that it turns out they will need), and other times it is better to be treated as *more* healthy (to avoid unnecessary treatments, or to qualify for higher-risk surgeries).[88] But any time the distribution of scores differs by race, users of scoring system will have to draw lines that places more error of one sort on the more vulnerable group.

The practice of thinking through and justifying the assigned bias-treatment options in advance may render the chosen option legally defensible after the fact. This might resemble the way that automobile manufacturers can avoid products liability even when confronted with a plaintiff's alternative design by showing that the plaintiff's preferred design is safer for some types of accidents but

---

[86] *See* discussion *supra* Section III.A.3.

[87] Darshali A. Vyas, Leo G. Eisenstein, & David S. Jones, *Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms*, 383 NEW ENGL. J. MED. 874, 876-77 (2020)

[88] *Id.* And in one case, the algorithmic scoring system seemed to be needlessly flawed by failing to account for important non-race factors. *Id.* (discussing the American Heart Association guidelines which assigned three additional risk points to any patient identified as non-Black).

less safe for other, more common ones.[89]

C. Biased Objective Function

Scholars often warn against strong presumptions that computer-aided decision-making will remove bias because there are multiple points in the algorithm-design and -training process where human error, past discrimination, and other social factors can bake bias into the algorithm.[90] The most consequential decision point, in our opinion, is during the selection of the *objective function*.[91] AI developers have to decide what output the algorithm should be forecasting.

AI scoring systems are always trying to predict something that is important to the decision-maker but that has not yet been observed. This is the *ultimate goal.* For example, going back to the example with the math test prediction, the eventual score on a math test might be the ultimate goal that the algorithm was trying to estimate. If this is so, it will be fairly easy to choose the objective function. The objective function is future performance on a specific math test, and the algorithm's performance can later be judged and the model improved based on that future performance. To train the algorithm, the programmers might have had training data that included many input variables and *outcome labels*—outcomes for students that took the same or a similar math test. The training outcome labels, the objective function, and the ultimate goal are all

---

[89] Aaron D. Twerski & James A. Henderson, Jr., *Manufacturers' Liability for Defective Product Designs: The Triumph of Risk-Utility*, 74 BROOKLYN L. REV. 1061, 1079-93 (2009). Note that the rules of product liability are stricter than the negligence rules that usually apply to "services" like algorithms and decision-making processes. For a further discussion of these issues, see generally Catherine M. Sharkey, *Products Liability in the Digital Age: Online Platforms as "Cheapest Cost Avoiders*, 73 HASTINGS L.J. 1327 (2022).

[90] Dan L. Burk, *Algorithmic Legal Metrics,* 96 NOTRE DAME L. REV. 1147, 1160 (2021) ("Algorithmic pattern detection and scoring outputs are not found, they are actually *constructed* by the processes of data harvesting, ingestion, and analysis.").

[91] Joel Shapiro, *Why Objective Functions Matter More as Companies Pivot in 2021,* FORBES (Dec. 29, 2020, 12:23 PM EST), https://www.forbes.com/sites/joelshapiro/2021/12/29/why-objective-functions-matter-more-as-companies-pivot-in-2021/?sh=54d16396531a [https://perma.cc/3HK6-64T6]; *see* Andrew D. Selbst, Suresh Venkatasubramanian & I. Elizabeth Kumar, *Deconstructing Design Decisions: Why Courts Must Interrogate Machine Learning and Other Technologies,* 85 OHIO ST. L.J. 415, 427-29 (2024).

very similar and thus easy to align: they are all (past and future) math test scores.

However, it's highly unlikely that performance on a math test is a valuable piece of information for its own sake. Suppose instead what the users of a math-performance algorithm *really* care about is future competence and creativity in a science or engineering career. This ultimate goal is much harder to observe—it would take years for evidence of success to ever materialize, and even then, it would rarely be captured and made accessible as research data. Thus, scores on a future math test in this scenario are an objective function that serves as a proxy for the ultimate goal that everyone truly cares about. Put this way, one can see and immediately understand the huge gap between the objective function (performance on a test) and the ultimate goal (eventual interest and career success in STEM). There are many waypoints between the two—classes, degrees, friendships, teachers, hardships, financial factors, and random luck—that would undermine confidence in the objective function. This would suggest that inferences drawn from a math test, let alone from a prediction of how somebody would perform on a math test *if they had taken it*, should be taken with a grain of salt.

In other cases, the connection between the objective function and the ultimate goal are even more attenuated because the objective function selected is or was under the influence of social factors and human behavior. If an algorithm is trained to optimize the prediction of an outcome that is itself the product of human decision-making (and therefore human error and bias), the algorithm itself will become very good not at forecasting the ultimate goal (the unobservable characteristic that really matters) but an output that reflects historical human decisions.

Many of the examples cited in critical works such as Cathy O'Neil's *Weapons of Math Destruction* illustrate this problem. When a university trains an algorithm to predict which admissions applicants are most likely to be selected based on past data, the machines will replicate whatever biases were held by the human admissions teams that preceded it.[92] When recidivism risk-scoring

[92] *See* Lilah Burke, *The Death and Life of an Admissions Algorithm*, INSIDE HIGHER EDUC. (Dec. 13, 2020), https://www.insidehighered.com/admissions/article/2020/12/14/u-texas-will-

algorithms use subsequent arrest as the objective function (rather than probability of committing a crime, which would be the ultimate goal of a recidivism risk score), the myriad social factors that affect when and where law enforcement make arrests will be imbedded in the model.[93] And if an algorithm uses historical home prices or search queries to estimate a home value or a query completion (respectively), the algorithm will make its predictions based on how humans *have* valued houses or used a search engine, and not necessarily based on how they *will* or *should* behave.[94]

In some (but not all) cases, the bias introduced from the wrong objective function can be reduced by selecting a different output for optimization. A better objective function will have a closer and less-contaminated relationship to the unobservable ultimate goal that its users are really interested in. Consider algorithms used for university-admissions purposes. A school has several options to train a machine-learning algorithm to predict what the school is truly interested in. Even if the school lacks good data on how past applicants or admitted students fared ten or twenty years later, and even if there is no general consensus on what it means for a college graduate to truly "succeed," a school could still select an output that is closer to that elusive ideal than past admissions decisions would be. For example, a college certainly has good data on whether admitted students completed a degree, earned good grades, and exceled or at least persisted in their selected major (in some instances information on subsequent employment might be available as well). All these outputs, though not entirely free from social influence, are more objective and, presumably, better linked to long-term goals of higher education than simply replicating the past admissions decisions of human administrators. Moreover, nothing prevents a university from choosing a more complex output

---

stop-using-controversial-algorithm-evaluate-phd [https://perma.cc/S7Z3-AMUW] (reporting the recission of an admission algorithm over concerns that it had incorporated human biases).

[93] *See, e.g.*, NORTHPOINTE, INC., PRACTITIONERS GUIDE TO COMPAS CORE 24-25, 25 tbl.3.8 (2015), https://archive.epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-COMPASPractionerGuide.pdf [https://perma.cc/65GZ-J38H] (summarizing the strength of correlations between various social factors and intake processing at Michigan Department of Corrections facilities).

[94] *See* SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM 167 (2018).

function to represent their complex, multi-objective goals. Perhaps the outcome that best represents the university's goals incorporates extra weight for diverse interests or for overcoming disadvantage.[95] A multicriteria outcome measure can be constructed, and probably should at any time an organization makes a highly consequential decision based on a complex notion of merit.

Several firms in the private sector have taken the lead on this issue. Companies like Zillow (a digital listing platform for real estate) and Google have altered their objective function to predict the outcomes that better match collective social goals.[96] In Zillow's case, recent external research shows that Zillow's "Zestimate" could make more accurate predictions of ultimate selling prices by incorporating the racial makeup of a neighborhood.[97] By including that information as a possible input variable, the Zestimate would be closer to the actual eventual selling price. But the company (presumably intentionally) has constrained the model by excluding the use of racial demographics as a factor in its prediction function.[98] In other words, Zillow seems to have decided to estimate not what the house is most likely to sell for, but what the house would be most likely to sell for *if buyers did not care about the racial makeup of its neighborhood*. The latter is arguably closer to the unobservable ultimate goal (what a home is really "worth" in some platonic sense) than the actual short-term selling price of the home.[99]

The same researchers found that Zillow's "Zestimates" had the effect, over time, of *changing* the prices at which homes were sold

---

[95] One type of diversity-aware objective function is discussed in Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, & Richard Zemel, *Fairness Through Awareness*, ITCS '12: PROC. 3D INNOVATIONS IN THEORETICAL COMP. SCI. CONF. 214, 224 (2012).

[96] In Google's case, the company has programmed the autofill generator to avoid showing disparaging results for individuals or certain demographic groups. *How Google Autocomplete Predictions Work*, GOOGLE, https://support.google.com/websearch/answer/7368877?hl=en [https://perma.cc/L9SJ-CKJW].

[97] Shuyi Yu, Digital Technologies, Customer Experience, and Decisions 43-44 (Mar. 10, 2021) (Ph.D. Dissertation, Massachusetts Institute of Technology), https://dspace.mit.edu/bitstream/handle/1721.1/139170/Yu-shuyiyu-PhD-Sloan-2021-thesis.pdf [https://perma.cc/G5KZ-S37Z].

[98] *Id.* at 43-46.

[99] This illustration can be treated as a debiased objective function (as it is here) or as a successfully constrained model, in contrast to the unsuccessfully constrained examples described *supra* in Section III.A.2.

so that short-term selling prices have started to converge with race-neutral assessments of home value. This example further emphasizes the importance of fairness in algorithms that not only reflect reality, but shape it as well.

But there will always be cases where the available data inevitably leaves the imprint of past human decision-making. In the case of recidivism risk scores like COMPAS, the objective function used to train and assess the models considers rearrest rates over a certain number of years.[100] A scoring algorithm would ideally have access not to which arrestees are subsequently rearrested for some new crime, but which arrestees *actually commit* a new crime. Yet there is no source for this unbiased output measure. Outside a few narrow contexts where non-law-enforcement surveillance is prevalent, arrests or convictions will be the best measures of crime commission, imperfect as they are.

This does not relieve the AI ethics officer from exercising diligence. To the contrary, an ethics officer who knows that their outcome function is a noisy, human-influenced proxy of the ultimate goal of interest should understand that they must manage this flaw and take account of it when making other trade-offs and usage decisions or while making appropriate changes to the data or the algorithm itself. This issue, in fact is one that has always been hiding in clear sight. Discussions of algorithmic decisions have now required policymakers to reconsider their objective functions and the implications of their selection.

D. Disparate Impact Without Biased Error or Treatment

Finally, some conceive of bias as an inequality in treatment across groups without regard to whether any of the problems described above have occurred. Some scholars refer to this as a failure to achieve "demographic parity."[101] This type of bias is detected any time the distribution of scores for one group deviates significantly from the distribution of scores for another, even if output and treatment error are in perfect parity. This is the most demanding form of antibias fairness because the goal would, by design, conflict with versions of fairness that require parity of

---

[100] Lee et al., *supra* note 28.

[101] Sandra Wachter, Brent Mittelstadt & Chris Russell, *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI*, COMPUT. L. & SEC. REV., July 2021, at 1, 22.

treatment to individuals who have identical predictions.[102] This definition of bias would also conflict with existing civil rights laws.[103] After all, the Supreme Court has already found that an employer that abandons a planned testing program for employee promotion on account of its disparate impact on minority candidates has engaged in *disparate treatment* against the other groups, in violation of the civil rights of candidates who stood to benefit from the test.[104] The employer can only backtrack in this way if it can prove it had a strong basis in evidence to believe the test would have violated antidiscrimination law.[105] This would require the employer to show not only that the test produced disparities, but that it also served no business-related purpose—that is, had no nexus to the relevant job duties. Any government use of algorithms or AI that are tuned to avoid disparate impact from a system that is unbiased in its error could be found to be unconstitutional under the Equal Protection Clause for similar reasons.[106]

Even apart from the legal risks, correcting for disparate impacts regardless of error could come at significant cost to the overall accuracy of the system, and will often generate greater group differences in error rates (the other forms of bias) not only for the majority or presumptively dominant group, but for other disadvantaged groups. For example, if the criminal justice system required equal proportions of Black and white arrestees to be

---

[102] This is sometimes referred to as the "single-threshold rule" because it requires that the same score threshold be used to make a decision regardless of race, gender, class, or other demographic category. Huq, *supra* note 80, at 1116.

[103] Ricci v. DeStefano, 557 U.S. 557, 585 (2009) (finding that a fire department violated Title VII of the Civil Rights Act when it abandoned a promotion program after discovering that the test it had intended to use would have led to the promotion of disproportionate number of white and Hispanic candidates).

[104] *Id.* at 579 ("Our analysis begins with this premise: The City's actions would violate the disparate-treatment prohibition of Title VII absent some valid defense.")

[105] *Id.* at 563 ("We conclude that race-based action like the City's in this case is impermissible under Title VII unless the employer can demonstrate a strong basis in evidence that, had it not taken the action, it would have been liable under the disparate-impact statute."); *id.* at 582 (quoting Richmond v. J. A. Croson Co., 488 U. S. 469, 500 (1989)).

[106] Indeed, this train of logic weaves through the passages of *Students for Fair Admissions, Inc. v. Harvard*, 600 U.S. 181, 213-14, 215-19 (2023) (expressing disfavor of the use of race and rejecting general concern about drop-offs in representation as a sufficient justification).

detained before trial, the available evidence suggests that a greater proportion of released Black arrestees would subsequently be arrested for crimes than the proportion of white arrestees who do, and that this gap in recidivism would have been predictable.[107]

Nevertheless, inspired in part by the EEOC's "four-fifths" rule that treats deviations greater than four-fifths as presumptive indicators of disparate impact,[108] some computer scientists have recommended programming machine-learning algorithms to detect[109] and even automatically correct for[110] disparate impacts. This approach is gaining traction because it attempts to proactively correct for the effects of historical racism and bigotry. After all, Black and white individuals who look the same based on the observables in a database are likely to differ in terms of the hardship, social seclusion, or opportunities they had prior to the moment at which an AI is used and judgment is passed.[111]

This also appears to be a possible implication for one of the illustrations in the White House *Blueprint for an AI Bill of Rights*:

> A predictive model marketed as being able to predict whether students are likely to drop out of school was used by more than 500 universities across the country. The model was found to use race directly as a predictor, and also shown to have large disparities by race; Black students were as many as four times

---

[107] *Cf.* John Gramlich, *Black Imprisonment Rate in the U.S. Has Fallen by a Third Since 2006*, PEW RSCH. CTR. (May 6, 2020), https://www.pewresearch.org/short-reads/2020/05/06/share-of-black-white-hispanic-americans-in-prison-2018-vs-2006/ [https://perma.cc/G9PJ-Z7DQ] (reporting that Black Americans still make up a higher proportion of prisoners than white Americans).

[108] 29 CFR § 1607.4(D) (2025).

[109] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, &Suresh Venkatasubramanian, *Certifying and Removing Disparate Impact*, KDD '15: PROC. 21ST ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE, DISCOVERY & DATA MINING 259, 260 (2015) (using a "confusion matrix" to identify potential disparate impact, and then calculating a disparate impact/utility tradeoff).

[110] Dwork et al., *supra* note 93, at 215 (offering a protocol that ensures the demographics of the set of individuals receiving a classification are the same as the demographics of the underlying population (within a factor) while treating similar individuals as similarly as possible; the authors call this approach "fair affirmative action").

[111] *See* Issa Kohler-Hausmann, *What's the Point of Parity? Harvard, Groupness, and the Equal Protection Clause*, 115 NW. U. L. REV. ONLINE 1, 15-17 (2020).

> as likely as their otherwise similar white peers to be deemed at high risk of dropping out. These risk scores are used by advisors to guide students towards or away from majors, and some worry that they are being used to guide Black students away from math and science subjects.[112]

The illustration points to two problems with the predictive model that we have already discussed: it may use race where other data more relevant to study habits, course selection, or living conditions that would make the model both more accurate and less biased in its error.[113] Indeed, even if race improved the predictive accuracy of the model, its inclusion could still be morally flawed—especially from a deontological perspective.[114]

In addition, the example suggests that the problem may lay in the treatment consequences of an otherwise-unobjectionable model. If the predictions are equally accurate across groups, the treatment may still be biased if students with higher risk scores are disserved by their guidance counselors—if the counselors deter students from pursuing studies in math or science, for example. This is consistent with problems described earlier in Section III.B.2. But a third possibility, and perhaps the most natural reading of the paragraph, is that Black students should not be steered away from math and science *even if* the predictive model accurately estimated that the drop-out rate for those students could be reduced.

Thus, whatever the current state of American law, the proper response to disparate treatment is contested in the AI ethics community.[115] It is unclear, however, whether the use of AI will lead to a graver problem of disparate impact than the results that would unfold in a human-driven decision-making scheme. Thus, opting to a fully human process need not be considered as a relevant remedy to this concern. We turn to a broader discussion of this lingering comparative question now.

---

[112] *Blueprint for an AI Bill of Rights*, *supra* note 1, at 24.
[113] *See* discussion *supra* Section III.A.2.
[114] Tal Z. Zarsky, *An Analytic Challenge: Discrimination Theory in the Age of Predictive Analytics*, 14 I/S: J.L. & POL'Y INFO. SOC'Y 11, 18-19 (2017); *see generally* BENJAMIN EIDELSON, DISCRIMINATION AND DISRESPECT (2015) (discussing the special wrongfulness of discrimination).
[115] Lee et al., *supra* note 28 (reporting out uncertainty and a variety of sentiments among a roundtable of AI ethics scholars and practitioners).

E. Compared to What?

Accusations that a machine learning algorithm is biased in some way often lead to calls to reject automated scoring or decision-making altogether and to restore human systems of judgment.[116] This happened, for example, at the University of Texas Department of Computer Science when the admissions scoring algorithm was criticized for being trained on past admissions decisions and for not taking sufficient account of applicants' personal background statements.[117] Yet without thoughtful analysis of the forms of bias most pernicious in each context, criticisms of AI can unwittingly *increase* bias by increasing society's suspicion and rejection of automated algorithms in favor of fallible human judges. For this reason, NIST's *AI Risk Management Framework* recommends judging AI risks and performance against a human baseline.[118]

Consider again the much-maligned recidivism risk scoring algorithms like COMPAS. The few studies that compare the effects of using recidivism scores to bail and sentencing decisions made in their absence find that jailing is reduced for members of *every* race, and that pretrial detentions could be even further reduced if judges are removed from the decision-making.[119] These findings are at odds with the way COMPAS scores are portrayed in the popular media, but they are consistent with studies in other areas finding that machine algorithms, as actually implemented, tend to improve race and gender disparities rather than exacerbate them. One study modeled that a hiring algorithm would have had a positive effect on racial minority applicants compared to the status-quo recruiting

---

[116] Aziz Huq, *A Right to Human Decision*, 106 VA. L. REV. 611, 620-28 (2020) (describing the legal and policy arguments that reject automation in decision-making, and challenging the reasoning of these arguments).

[117] Burke, *supra* note 90.

[118] NIST AI RISK MANAGEMENT FRAMEWORK, *supra* note 6, at 6.

[119] JOHN KLEINBERG, HIMABINDU LAKKARAJU, JURE LESKOVEC, JENS LUDWIG & SENDHIL MULLAINATHAN, *HUMAN DECISIONS AND MACHINE PREDICTIONS*, 133 Q.J. ECON. 237, 270-72, 275-78 (2018); *SEE ALSO* MEGAN STEVENSON, *ASSESSING RISK ASSESSMENT IN ACTION,* 103 Minn. L. Rev. 303, 356, 361, 368-69 (2018) (finding that a law requiring judges to at least consider risk assessment scores caused a short-term reduction in pretrial detention, but that the reduction faded over time as judges returned to their previous habits; finding that pretrial arrests increased when the scores were influencing judge's decisions, but that pretrial arrests for violent crimes went down slightly; but not finding promising reductions in the race gap).

process,[120] and another found that a machine learning algorithm could be used to select corporate directors who were more likely to be female *and* more likely to outperform the directors selected by the company board.[121] Also, home mortgages tend to have lower interest rates and lower default rates when banks make use of Big Data profiles that go beyond the income and credit-score information typically collected, suggesting that machine learning has promise for helping low-income applicants prove that they are more creditworthy than loan officers have historically thought.[122]

So, while theoretical accounts of the ways that AI can exacerbate bias are sound, the empirical studies that attempt to directly compare machine-learning performance to its purely human alternative suggest automated algorithms in practice tend to *reduce* bias instead of increasing it (and often increase accuracy at the same time).[123]

This is not to say that AI and machine-learning algorithms should be presumed to be ethical. They should be designed, audited and held accountable to a standard that is practical and reasonable given the available data and technology, and law or industry norms should push for progress along all the dimensions described in this

---

[120] Bo Cowgill, *Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening* 27-30 (Upjohn Inst. Working Paper, Paper No. 19-309), https://ssrn.com/abstract=3584916 [https://perma.cc/6QBH-WGKE].

[121] ISIL EREL, LÉA H. STERN, CHENHAO TAN & MICHAEL S. WEISBACH, *SELECTING DIRECTORS USING MACHINE LEARNING*, 34 REV. FIN. STUD. 3226, 3229 (2021).

[122] Jin-Hyuk Kim & Liad Wagman, *Screening Incentives and Privacy Protection in Financial Markets: A Theoretical and Empirical Analysis*, 46 RAND J. ECON. 1, 17-18 (2015); *see also* Will Dobie, Andres Liberman, Daniel Paravisini & Vikram Pathania, *Measuring Bias in Consumer Lending*, 88 REV. ECON. STUD. 2799, 2823-28 (2021) (demonstrating that consumer-lending decisions made using machine-learning predictions of long-run profits both increased profits and eliminated bias, compared to human loan examiners). By contrast, studies of human-driven mortgage lending decisions continue to find racial bias even after controlling for credit history and income. *Id.* at 2799 (citing Kerwin Kofi Charles, Erik Hurst & Melvin Stephens, *Rates for Vehicle Loans: Race and Loan Source*, 98 AM. ECON. REV. 315 (2008); and Patrick Bayer, Fernando Ferreira & Stephen L. Ross, *What Drives Racial and Ethnic Differences in High-Cost Mortgages? The Role of High-Risk Lenders*, 31 REV. FIN. STUD. 175 (2017).

[123] In addition to the studies described *supra* notes 118-120, see also Omri Ben Shachar*, Exploring the Regulatory Resistance to Data Technology in Auto Insurance,* 15 J. LEGAL ANALYSIS 129, 136-143 (2023), which shows that intensive data-collection by auto insurance helps educate drivers so that they reduce both accidents *and* insurance payments.

taxonomy. Our point is only that resort to human systems is not necessarily better—and might be worse. Given the ability of machine systems to substantially reduce noise and counterproductive biases, two known flaws in human systems,[124] we should expect the standards for AI and machine learning systems to eventually surpass the crude antidiscrimination rules that are in place today. This is in addition to the fact that these tools can account for and integrate all relevant factors to the final decision.

IV. Unfairness as Disproportionality

If an automated algorithm meets a data steward's standards for accuracy and equity, it could still be *perceived* as unfair if it gives great weight to a seemingly arbitrary or insignificant factor. Consider as an example a 2008 FTC inquiry into the analytical practices of a bank and credit-card issuer that limited the credit limit of its customers based on usage of the credit card at certain types of businesses including pawn shops, massage parlors, counseling services, or billiard halls.[125]

These credit practices seemed intuitively unacceptable, in part because of the privacy harm that comes from repurposing data collected for one purpose to serve another without adequate customer consent (although the firms reserved the rights to do so in the issuing agreement all consumers signed, something consumers probably did not notice).[126] Another explanation for the concern is that there may be injustice when a small change in an input variable results in a large difference in how a person is treated. Even if historical data suggests that a person who frequents a billiard hall really is on a different trajectory, without a causal theory, it seems

---

[124] KAHNEMAN ET AL., *supra* note 39, at 6.

[125] Complaint at 7-8, 34-35, FTC v. CompuCredit Corp., No. 1:08-cv-01976-BBM (N.D. Ga. June 10, 2008). The FTC's investigation led to a settlement with the firms, according to which the firms undertook to provide proper and specific disclosures of such credit restriction factors prior to relying on them. Press Release, FTC, Subprime Credit Card Marketer to Provide At Least $114 Million in Consumer Redress to Settle FTC Charges of Deceptive Conduct (Dec. 19, 2008), https://www.ftc.gov/news-events/news/press-releases/2008/12/subprime-credit-card-marketer-provide-least-114-million-consumer-redress-settle-ftc-charges [https://perma.cc/CZH9-X4BJ].

[126] This is the notion of "purpose limitation" in EU data-protection law. GDPR, *supra* note 29, at 35.

arbitrary and disproportionate to treat that person significantly worse than another who is identical in every other way based on this somewhat trivial distinction. Nonetheless, such an outcome might result from an algorithmic design that is working as intended. This is a well-documented phenomenon in machine learning which often creates a variety of "bins" that make seemingly meaningless distinctions between people but create sharp discontinuities in the output predictions.[127] This technological feature explains why machine learning can exacerbate both the perceived and actual problem of steep cutoff points which exist in any regulatory system.[128]

Although there is not yet a robust discussion of the bounds and value of this form of unfairness, the intuition matches notions of proportionality and requirements to be rational and nonarbitrary in the law.[129] For example, in guidance issued in the EEOC's December 2000 Compliance Manual, the EEOC states that "the difference in education, experience, training, or ability must correspond to the compensation disparity. Thus, a very slight difference in experience would not justify a significant compensation disparity."[130] Courts examining this issue were required to establish whether a substantial difference in pay can be justified by teaching in different university departments, for example. We therefore predict to see increased engagement with the technical questions that these sorts of policies raise.

Drawing a clear normative conclusion on proportionality is difficult. So, as with other forms of unfairness, the "fair enough" principle should hold. A model that produces a "small change, big difference" dynamic should raise questions and follow-up testing to see if the dynamic can be reduced without reducing accuracy, increasing bias, or compromising other forms of fairness. If it can, it should. If it cannot, the AI tool will have to operate in the land of tradeoffs. Disproportionality may be legally and ethically tolerable as long as it is not egregiously problematic regarding the other aspects discussed.

---

[127] Bambauer et al., *supra* note 40, at 2367.
[128] *Id.* at 2396.
[129] *Id.* at 2350.
[130] U.S. EQUAL EMP. OPPORTUNITY COMM'N, SECTION 10 COMPENSATION DISCRIMINATION (2000).

V. Unfairness as Manipulability

Another somewhat underexplored aspect of fairness is whether an algorithmic decision-making process induces excessive strategic behavior, or "gaming."[131] Complex decision-making models can give an advantage to individuals who have better information about the variables that matter most, especially if these are changeable (or not immutable). To use an example from the "little data" world, better-educated Americans know much more about the factors used to produce FICO scores—one of the most prominent pieces of data that banks use to make decisions on lending. The privileged understand that it is better to have multiple lines of credit, and to use them and pay them regularly, while an equally responsible person from a less privileged background would not see the value in opening multiple credit cards.[132] We refer to this as a "little data" example because the FICO score algorithm involves the analysis of several recognized variables to create a score, as opposed to applying more-complex models, featuring "nontraditional data" used by fintech firms.[133]

Manipulability of an algorithm is problematic for many reasons. First, it can be exploited by data subjects or by hackers engaged in "adversarial machine learning" so that resources are allocated inaccurately, and possibly in favor of people who already have access to more resources (like information and technical ability).[134]

---

[131] For a complete treatment of the manipulability problem, *see generally* Jane Bambauer & Tal Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV. 1 (2018). Joshua Kroll and his coauthors have acknowledged that data subjects can engage in strategic behavior that could render algorithm transparency undesirable even if it were possible (which they doubt). Joshua Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633, 639 (2017).

[132] *See generally* Annamaria Lusardi & Jialu L. Streeter, *Financial Literacy and Financial Well-Being: Evidence from the U.S.*, 1 J. FIN. LITERACY & WELLBEING. 169 (2023) (discussing disparities in financial literacy based on socioeconomic factors).

[133] Majid Bazarbash, *FinTech in Financial Inclusion Machine Learning Applications in Assessing Credit Risk* 26 (Int'l Monetary Fund Working Paper, Working Paper No. 19/109, 2019), https://ssrn.com/abstract=3404066 [https://perma.cc/F5HS-QH67].

[134] Bambauer & Zarsky, *supra* note 129, at 11-12.

In this sense, manipulability is a species of the accuracy and bias problems discussed above. But gaming also introduces a distinct form of unfairness. By being manipulable, an algorithm imposes a burden on data subjects to self-monitor and constantly assess whether they want to make certain superficial behavioral changes in order to get a better score—especially when the factors considered are mutable. Manipulability therefore captures the anxiety and self-censorship that general surveillance can create.

Decision-making processes were always somewhat subject to manipulation. The move from human to automated decisions, and later from small- to big-data processes did not necessarily generate more manipulation options—just different ones. It shifted the power from the usual suspects who have learned how to tame systems to their preferences to a different set of successful manipulators (some old, some new). In addition, the specter of an automated, data-driven, computerized process might lead many to assume that the system is beyond tampering. While this presumption is clearly false, the misunderstanding affects both the motivation to manipulate, and the consequences to people subject to manipulated systems. As with other forms of fairness, manipulability cannot be reduced to zero, and almost any reduction will likely impact other forms of fairness. Thus, again, we recommend reducing manipulability to the extent that it is a needless flaw, and otherwise to embrace a "fair enough" approach by allowing some flexibility to make tradeoffs consciously and thoughtfully.

VI. Unfairness as Opaqueness

AI typically features highly complex and ever-changing decision models that are not amenable to transparency in the traditional sense.[135] When a highly consequential decision is made about an individual, there is commonly a cultural expectation (and sometimes a legal requirement) that the individual receive some explanation for the decision, especially if it is unfavorable to them.[136] When the government takes adverse action or exercises its

---

[135] *See generally* FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015) (discussing the secrecy around algorithms used by both public and private entities).

[136] Goldberg v. Kelly, 397 U.S. 254, 267-68 (1970) (requiring the government to provide notice of the reasons for a termination of government benefits).

administrative powers, due process requires some means to understand and challenge the decision.[137] The concept is also imbedded in the American criminal justice system through the right to confront and question one's accusers and the necessity for police to have articulable suspicion to justify an arrest.[138]

Some have argued that black-box algorithms are illegitimate based on their opaqueness alone since explanation has historically been a necessary component of accountability in human systems.[139] Receiving an explanation might also be seen as an important extension of individuals' autonomy and control over personal data pertaining to them. But the discussion of transparency in AI systems has evolved in recent years to value accountability over descriptions and causal theories.[140] Accountability can take forms that do not require giving up or explaining source code and training data. It can take the form of "counterfactuals"—explaining how much a use must "improve" or change in one dimension so that the applicant will quality for the relevant allocated good or service.[141] The algorithm can also be audited using test data to check for consistency, bias, or other problems.[142] And in addition to checking for unacceptable levels of unfairness (as defined in one or more of the ways already discussed), it can also operate a self-assessment of tradeoffs, quantifying how much of an improvement in one form of

---

[137] Benjamin Eidelson, *Reasoned Explanation and Political Accountability in the Roberts Court*, 130 YALE L. J. 1748, 1754 (2021) (describing the Due Process requirement for government to provide a reasoned explanation for a change of policy).

[138] Coy v. Iowa, 487 U.S. 1012, 1015-20 (1988) (recognizing a Sixth Amendment right to face-to-face confrontation); Terry v. Ohio, 392 U.S. 1, 21 (1968) (establishing that the Fourth Amendment requires officers to "point to specific and articulable facts" justifying a search or seizure).

[139] *See, e.g.*, PASQUALE, *supra* note 133, at 193. For a discussion of the "accountability" justification for transparency (as part of a broader set of transparency justifications) see Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503, 1533 (2013).

[140] *See, e.g.*, Brandon L. Garrett & Megan Stevenson, *Open Risk Assessment*, 38 BEHAV. SCI. & L. 279, 283 (2020) (noting deficiencies in the attempts at transparency because the reasoning for categorizing certain scores as "high" risk is not explained).

[141] Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH. 841, 882□83 (2018).

[142] *See, e.g.*, Aequitas, *Bias and Fairness Audit Toolkit*, GITHUB, https://github.com/dssg/aequitas [https://perma.cc/GGW2-RJ3K].

fairness will affect other forms of fairness.[143] Therefore, transparency (or lack thereof) need not be linked to more or less biases and discrimination.

Moreover, there are some advantages to opaqueness even in terms of fairness, as there is a natural tension between goals of increasing transparency and goals of reducing manipulability described in the last section. In addition, structuring algorithmic processes which enable transparency might compromise their accuracy—another competing notion of fairness. And alternative forms of accountability might also avoid conflict with legal rules protecting proprietary trade secrets. In any case, the losses in transparency might not be as great as they seem despite the opaqueness of machine learning practices. Humans, too, often engage in spurious sense making when called on for an explanation. The brain is a "black box" too.[144] Thus any alternative to the algorithmic process will feature opacity as well, yet again of a different flavor; while humans (subject to mandate) might share the reasons for their actions, such reasons might not be truthful.

Conclusion: Ubiquitous Unfairness

Given the cross-cutting goals and societal aspirations that affect how decision-making will be perceived, defining and creating a "fair" algorithm is primarily a policy task rather than a matter of technology or pure logic. This fact has been recognized in legal scholarship for some time.[145] The trouble is, recent AI regulatory

---

[143] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass Sunstein provide a good start. Kleinberg et al., *supra* note 2, at 149-50.

[144] Roland Hewage, *Deep Learning & Artificial Neural Networks: Solving the Black Box Mystery*, HACKERNOON (Apr. 9, 2020), https://hackernoon.com/deep-learning-and-artificial-neural-networks-solving-the-black-box-mystery-rl7h32wa [https://perma.cc/LST8-WQX8]; *see generally* KEVIN SIMLER & ROBIN HANSON, THE ELEPHANT IN THE BRAIN: HIDDEN MOTIVES IN EVERYDAY LIFE (2018) (discussing hidden, subconscious, and self-deceptive motives for human behavior).

[145] Sandra Mayson's seminal article *Bias In, Bias Out* gave a very clear and balanced assessment of the policy tradeoffs that must be made. Mayson, *supra* note 5, at 2248-49; *see also* Richard Berk, *Accuracy and Fairness for Juvenile Justice Risk Assessments*, 16 J. EMP. LEG. STUD. 175, 175 (2019) ("Although statisticians and computer scientists can document the tradeoffs, they cannot

frameworks have demonstrated an unwillingness to state which types of unfairness will be tolerated in order to avoid other forms of unfairness.[146] Implementing one measure to promote fairness might at times generate or exacerbate fairness on another dimension. We suspect that vagueness and abdication of decision-making will plague AI public policy debates for the foreseeable future. Setting priorities not only raises disagreements between regulators, but also creates headaches for each individual lawmaker, too, who must answer to media inquiries, firms, and voters armed with examples of bias, opaqueness, inaccuracy, and privacy intrusions, no matter what option the lawmaker chooses. The public is not prepared for a frank admission that it is acceptable for a large AI company to decide, in advance, that it is acceptable to implement an algorithm that will be wrong more often for one group than another. Nor is it prepared to hear that the same company decided in advance to reduce accuracy for everybody to relieve some forms of bias (but not all).

This Article attempts to steer industry norms and public debate toward a style of analysis that can cut through the negative rhetoric. Rather than focusing on what is lost as compared to a perfect baseline, firms and regulators should begin to assess AI ethics through a frame of growth and marginal improvements. In particular, we suggest that during this critical time of development and commercialization, industry participants should choose their tradeoffs between the various dimensions of fairness consciously. In the absence of clear legal instructions that prioritize some forms of fairness over others, regulators should tolerate some variance in how the inevitable tradeoffs are made. The discussion in this article has some additional implications for public policy:

AI ethics should apply to human systems as well. Human

---

provide technical solutions that satisfy all fairness and accuracy objectives. In the end, it falls to stakeholders to do the required balancing using legal and legislative procedures, just as it always has.").

[146] *See supra* text accompanying note 1. The NIST Risk Management Framework is alone among policy guidance with high visibility in acknowledging this. "Addressing AI trustworthiness characteristics individually will not ensure AI system trustworthiness; tradeoffs are usually involved, rarely do all characteristics apply in every setting, and some will be more or less important in any given situation. Ultimately, trustworthiness is a social concept that ranges across a spectrum and is only as strong as its weakest characteristics." NIST AI RISK MANAGEMENT FRAMEWORK, *supra* note 6, at 12.

systems of prediction and judgment implicate every one of the ethical issues discussed in this article. However, human systems are so disaggregated, disorganized, and constrained that hierarchies of values are less salient (because they are scattered and hard to observe) and less relevant (because improvements along any dimension of fairness are difficult).[147] Thus, proposals to severely limit or ban AI decision-making for fairness reasons should first ensure that the existing system does not exhibit the problems motivating the ban.[148] As noted above, almost every one of the problems catalogued here is a feature of human decision-making as well.

Some charges of unfairness are more valid than others. An accusation that an algorithm is inaccurate, biased, overly opaque, or too gameable will be valid if the faults are unnecessary—if they are known or reasonably discoverable and can be corrected without significantly degrading other forms of fairness. Thus, while we have emphasized that ethical tradeoffs must be made during AI design, that is only true for applications and designs that have already made every Pareto-efficient improvement.[149] If an AI application needlessly compromises accuracy, bias, or some other aspect of fairness, it deserves criticism. Any time a company can make improvements for minimal costs along the other dimensions of fairness, they should. The most worrying criticisms are those that are made without any attempt to assess whether the perceived problem is easy to fix (without tradeoffs) or is difficult, requiring compromise between values.

Industry experimentation will shape debate and can make genuine progress. This Article presented a concise and implementable taxonomy of fairness. If AI companies assess their programs along these dimensions as best they practicably can, and then document why certain design choices were made or retained, they can convert a vague regulatory self-assessment process into an instructive and genuinely valuable articulation of fairness in the

---

[147] KAHNEMAN ET AL., *supra* note 39, at 34-38 (describing noise in "singular decisions" where individuals don't see similar facts more than once).

[148] California's AI inventory at least contemplates that alternatives to the automated decision system are explicitly listed and that the results of "any research assessing the efficacy and relative benefits of the uses and alternatives of the automated decision systems" also be made public. CAL. GOV. CODE § 11546.45.5(c)(2) (West 2025).

[149] Xu & Strohmer, *supra* note 4, at 6.

context in which the algorithm will be used. This will spur learning and a tolerance for nuance among developers, users, regulators, and the public. To facilitate this, lawmakers and regulators can create legal sandboxes in which AI companies are shielded from legal risks in exchange for radical transparency. Utah has created something like this through a "learning lab" within its Office of Artificial Intelligence Policy.[150] The Learning Lab coordinates with private companies as they implement an AI application so that government and industry can generate best practices together.[151]

The challenges of achieving fairness in machine learning and AI-powered processes is not going away any time soon. Quite to the contrary, it is destined to invoke even harsher disagreements and problems. The growing use of GenAI tools (such as OpenAI's ChatGPT) will bring many more individuals to an interface that provides responses premised on a specific model of performance and fairness. They will experience the concern and frustration that comes with relying on models that are not calibrated to their own values and priorities. These developments will amplify demands for analytical frameworks that can balance, compare, and compromise between different visions of fairness. This Article provides a means to organize, assess, and tolerate a range of AI models that are all imperfect but fair-enough.

---

[150] Off. of A.I. Pol'y, *Learning Lab*, UTAH DEP'T OF COM., https://ai.utah.gov/learning-lab/ [https://perma.cc/KNV7-6MJH].

[151] Zoom Interview by Jane Bambauer, Brechner Eminent Scholar, Levin Coll. of L., with Brady Young, Lead A.I. Legal & Pol'y Analyst, Utah Dep't of Com. (Aug. 10, 2024).