The Fairness-Accuracy Tradeoff Myth in AI

Ignacio Cofone[*]

Draft for WeRobot 2025

Abstract

This paper examines an understated reason for which we should push back against the pervasive but flawed notion of a "fairness-accuracy tradeoff" in AI. The idea persists in academic, industry, and policy discourse, yet it rests on a fundamental misunderstanding of how AI models function and how they are used in society. AI outputs aren't direct indications of ground truth—they are proxies for decisions, meaning that "accuracy" is always relative to an arbitrarily chosen output variable. Because AI models optimize for imperfect output variables rather than ground truth, adjusting for fairness doesn't necessarily reduce accuracy; in many cases, it can enhance it. The real question isn't whether fairness comes at the cost of accuracy, but whose definition of accuracy is being prioritized and why. The assumption of a universal fairness-accuracy tradeoff is tempered significantly when examined through the lens of model multiplicity and varying fairness definitions, many of which include accuracy. But it collapses when examined in light of biases in the very output variables used to measure whether de-biasing reduces accuracy. Any claim of a tradeoff depends on what is being optimized—which output variable is chosen and whether it embeds the very biases fairness constraints seek to correct. An AI model that optimizes for a biased output variable may be "accurate" with respect to that flawed variable but inaccurate with respect to the underlying characteristic that actually matters for decision-making. Thus, there can be no general tradeoff between fairness and accuracy—only competing choices about which metrics to accept and whose interests to prioritize.

[*] Please forgive the unedited writing and unfinished footnotes. I've been giving this issue a good amount of thought for a while but still need to edit this draft significantly. I am very much looking forward to your critical comments and feedback. Please do contact me for an updated version if you'd like to use it or share it beyond participants.

[*] Professor of Law and Regulation of AI, University of Oxford, Faculty of Law and Institute for Ethics in AI; affiliate member, Quebec AI Institute. ignacio.cofone@law.ox.ac.uk.

1) Introduction

The often-stated claim that implementing fairness constraints in machine learning models inevitably leads to reduced predictive accuracy is a myth.[1] Building on research on algorithmic fairness and sociotechnical systems, this paper argues that there is no demonstrable general tradeoff between fairness and accuracy. In fact, often, such as in many situations of hiring and workforce management, fairer algorithms can enhance overall predictive performance over the right metrics. It can also lead to more generalizable models.

The idea that there is a necessary tradeoff between fairness and accuracy in AI stems from the assumption that improving fairness necessarily reduces a model's predictive performance. The tradeoff, which leads to the idea that promoting fairness in AI reduces its accuracy at an often hefty social cost, is popular in academia, policy circles, and the press.[2] It forms part of a set of broad technical claims that are relevant to the law in insidious ways.[3] Many argue, for example, that "minimizing discrimination will unavoidably involve a tradeoff with accuracy."[4] In one of its most common applications, employment, the tradeoff is presented as one between algorithmic fairness and hiring job candidates who are underqualified for a position, imposing costs on employers.

The reason for which the fairness-accuracy tradeoff claim, which formulates it generally for all AI models, is false is that it fails to consider an important point from sociotechnical literature: the output variables of AI models, such as the similarity of a job candidate to current employees, are themselves sitting for decision criteria, such as who will be a good employee. They form a bridge between the available data that AI can process and the intangible qualities that society values. This consideration adds to two bodies of literature. First, what the extensive literature about proxy bias explains of algorithmic labels being proxies for ground truth facts is also true of output variables being proxies for normative facts. Second, results from research on model multiplicity that show that fairness can be improved for models that don't achieve the highest possible accuracy (in terms of output variable) in their predictions given their dataset can be generalized further because one doesn't need to accept accuracy measures from output variables.[5]

This gap between models' objectives and social objectives is a problem for developing and deploying human-centered AI because social and private incentives to reduce it may differ significantly.[6] Framing algorithmic fairness as being in a general tradeoff with accuracy is a convenient theory for those who want to resist implementing AI fairness measures into their systems, but the tradeoff doesn't capture the whole picture.[7] The rhetoric of a fairness-accuracy tradeoff isn't a neutral analytical tool; it shields decision-makers seeking to deploy models that reinforce systemic disparities or are concerned only with cost-reduction with indifference towards systemic disparities, often under the guise of technological inevitability. Framing fairness and accuracy as opposing values obscures the well-known fact that AI doesn't discover ground truth—it optimizes for human-defined objectives. The relevant normative choices aren't just about which model and which labels to use, but also which objective to set, which output variables to prioritize, and, ultimately, whose version of accuracy governs the system.

[1] Here, I use accuracy as the proportion of correctly classified instances out of the total instances with regards to the model's prediction objective.

[2] See Section 2.a.

[3] Ryan Calo, "How we talk about AI (and why it matters)" (2019) *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* 1.

[4] Emily Black et al., *The Legal Duty to Search for Less Discriminatory Algorithms*, ACM 1, 1 (2024), https://arxiv.org/pdf/2406.06817 (presenting, not endorsing, the argument). See, e.g., https://5harad.com/papers/fairness.pdf

[5] Emily Black and others, 'The Legal Duty to Search for Less Discriminatory Algorithms', *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery 2024).

[6] See Ignacio Cofone & Katherine Strandburg, *Unjustifiable Algorithmic Opacity* (draft 2025) (explaining the principal-agent problem between algorithmic designers and society that allows for algorithmic decision-making).

[7] See Emily Black and others, 'Less Discriminatory Algorithms' (2024) 113 Georgetown Law Journal 1.

The legal and policy community should therefore move beyond the false dichotomy of fairness versus accuracy and instead scrutinize who benefits and who loses when fairness is prioritized in design choices.

The concern regarding the accuracy cost of implementing fairness in AI also overlooks the benefits of such measures. First, focusing solely on the accuracy of output variable values at the expense of fairness perpetuates and amplifies biases, leading to outcomes that aren't only normatively problematic but also less effective in real-world scenarios different from the one the model was trained on.[8] Second, fairness can enhance the reliability of Ai models, thereby expanding their utility and leading to more robust models, mitigating concerns over any impression of a dip in accuracy. Third, the argument that prioritizing fairness necessarily compromises overall accuracy is mistakenly static: it fails to account for iterative improvements in AI development, which mean that fairness measures that prioritize the prediction objective over the output variable can lead to more inclusive developments.

Section 2 unpacks the alleged tradeoff by exploring the statistical mechanics that exist behind AI discrimination and the conditions imposed by AI fairness. Section 3 argues that the general statement of the tradeoff is overstated by building on recent literature. Section 4 introduces an analysis from a sociotechnical systems perspective to show that it's false; it applies this argument to risk assessment algorithms to illustrate generalizability outside of the employment context. Section 5 examining the extent to which it matters for the law and for decision-making policy. Section 6 proposes a test to determine in which situations the tradeoff may take place.

2) The Tradeoff Claim

The tradeoff claim can be found in industry and government. Google's guidance on AI governance, for example, proposes the tradeoff in its 2020 AI regulation whitepaper: Google notes that different technical fairness approaches yield models equitable in different ways, "and may require tradeoffs in terms of general accuracy or efficiency," understanding that fairness can constrain a model's peak performance.[9] Practitioners have repeatedly stated this tradeoff. For example, an RMA Model Risk consortium notes that "enforcing a fairness criterion does come at the cost of accuracy," because fairness constraints act as additional restrictions on a model.[10] A 2023 DARPA program call claimed that there are "well-studied trade-offs between accuracy and fairness of AI models," stating "some degree of fairness may be sacrificed…to optimize for the accuracy" of a prediction.[11]

The media has several examples too.[12]

[8] See, e.g., Pauline Kim, 'Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action' (2022) 110 California Law Review 1539, 1548. ("Statistical bias can result when the data used to train the model are unrepresentative of the population or contain systematic errors")

[9] https://ai.google/static/documents/recommendations-for-regulating-ai.pdf

[10] https://kdoden.com/wp-content/uploads/2020/12/RMA-FAIR-ML-Session-2.pdf

[11] https://www.darpa.mil/research/programs/analyzing-trade-off-bias-accuracy

[12] See, e.g., Irineo Cabreros, 'Op-Ed: Why an Algorithm Can Never Truly Be "Fair"' *Los Angeles Times* (27 March 2022) <https://www.latimes.com/opinion/story/2022-03-27/algorithms-unfair-racial-bias-math>.

And the fairness accuracy tradeoff is alive and well in academia across disciplines: computer science,[13] management,[14] psychology,[15] and law.[16] These include a number of papers that come after the work on why the tradeoff is overstated.[17]

The argument often extends from *whether* to implement fairness to *how* to implement fairness. Fair AI measures can also be classified into two categories, which I will call "process-based" and "classification-based." Process-based fair algorithms transform training datasets to remove any dependency between a perceived attribute (e.g., race, gender) and a target attribute (e.g., hireability). When the data points are processed before being fed to a model, for example by re-sampling to satisfy fairness constraints,[18] it's called pre-processing.[19] Post-processing, on the other hand, involves processing the data after it has been fed to a model, which can be done by relabeling the data to train fair classifiers[20] or modifying it after the model is trained. [21] Classification-based fair algorithms regress data points in a fairness-adjusted way through the structure of the model itself. Some examples are a discrimination loss

[13] See, e.g., Han Zhao and Geoffrey J Gordon, 'Inherent Tradeoffs in Learning Fair Representations' (2022) 23 Journal of Machine Learning Research 1, 2 (presenting a tradeoff between statistical parity and accuracy); Irene Y. Chen, Fredrik D. Johansson & David Sontag, *Why is my Classifier Discriminatory?*, 31 Adv. in Neural Info. Process. Syst. 3539 (2018) (arguing that balancing accuracy with fairness leads to undesirable outcomes and unfairness should be addressed at data collection but not by constraining the model); AK Menon and RC Williamson, 'The cost of fairness in binary classification', *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (2018); Sam Corbett-Davies and others, 'Algorithmic Decision Making and the Cost of Fairness', *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery 2017) 800 (arguing that it's algorithms not subject to fairness constraints that maximize public safety).

[14] See, e.g., Mike Teodorescu and others, 'Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation' (2021) 45 Management Information Systems Quarterly 1483, 1486. ("Even in the most straightforward case of a protected attribute with two categories (e.g., gender), there is a tradeoff between accuracy, an oft-used metric for model performance, and fairness. [...] The model may need to sacrifice accuracy for the sake of a fair outcome."); Nathan Mondragon, 'Artificial Intelligence in Automated Scoring of Video Interviews' in Tracy Kantrowitz, Douglas H Reynolds and John Scott (eds), *Talent Assessment: Embracing Innovation and Mitigating Risk in the Digital Age* (Oxford University Press 2023) 111 ("tradeoffs are often involved. For example, [...] assessments with strong predictive validity such as general mental ability tests can have implications for adverse impact.").

[15] See, e.g., Caleb Rottman and others, 'New Strategies for Addressing the Diversity-Validity Dilemma with Big Data' (2023) 108 The Journal of Applied Psychology 1425, 1425. ("the diversity–validity dilemma, reflects the quandary where assessments with the highest predictive validity also tend to have the highest adverse impact [...]. This forces organizations to choose between hiring a diverse set of candidates or hiring the candidates most likely to succeed."); Keith M Pyburn, Robert E Ployhart and David A Kravitz, 'The Diversity–Validity Dilemma: Overview and Legal Context' (2008) 61 Personnel Psychology 143, 144. ("The ability of organizations to simultaneously identify high-quality candidates and establish a diverse work force can be hindered by the fact that many of the more predictive selection procedures negatively influence the pass rates of racioethnic minority group members (non-Whites) and women. This can create a diversity-validity dilemma").

[16] See, e.g., Sandra Mayson, *Bias in, bias out*, 128 Yale LJ 2218, 2298 (2018) ("equalizing false-positive and false-negative rates might increase the net burden of prediction on communities of color ... the greater total error rate overwhelms the greater per capita benefit"); Id. at 2300 ("if prioritizing equality in error rates imposes too great a cost in accuracy, it will eliminate the utility of prediction")

[17] See Section 3.

[18] Faisal Kamiran and Toon Calders, 'Classification with No Discrimination by Preferential Sampling', *Proceedings of the 19th Machine Learning Conference of Belgium and the Netherlands* (2010) 1–6. See also Black and others, 'Less Discriminatory Algorithms' (n 2) 39.

[19] See, e.g., Rich Zemel and others, 'Learning Fair Representations', *Proceedings of the 30th International Conference on Machine Learning* (PMLR 2013) <https://proceedings.mlr.press/v28/zemel13.html>.; Michael Feldman and others, 'Certifying and Removing Disparate Impact', *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery 2015) <https://doi.org/10.1145/2783258.2783311>.

[20] Indre Žliobaite, Faisal Kamiran and Toon Calders, 'Handling Conditional Discrimination', *Proceedings of the 2011 IEEE 11th International Conference on Data Mining* (IEEE Computer Society 2011) 992–1001 <https://doi.org/10.1109/ICDM.2011.72>.

[21] Moritz Hardt, Eric Price and Nathan Srebro, 'Equality of Opportunity in Supervised Learning', *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Curran Associates Inc 2016) 3324–30.

term,[22] shifting the decision boundary,[23] constrained optimization,[24] a regularizer,[25] and constrained optimization for classification.[26] Process-based fair algorithms are more generally applicable; because they work on the data instead of on the model itself, they can be paired with a large set of off-the-shelf AI classification and regression algorithms. They can also be cheaper than classification-based algorithms. The transformed data can be considered a fair representation that is free from discrimination.[27] But process-based algorithms are often critiqued on the basis that they reduce the accuracy of the output variable through information "loss."

However, as the next two sections, the accuracy concerns over process-based algorithms should be tempered. When one looks at the larger context, they aren't as concerning as they initially seem.

3) Why the Tradeoff is Overstated

a) Fairness Definitions

There are many definitions of AI fairness and they can't be satisfied at the same time.[28] Allow me to mention some of the most common ones to illustrate this: (a) demographic parity is satisfied when positive outcomes are proportionally distributed across demographic groups (e.g., hiring decisions should reflect the same selection rate for different racial or gender groups); (b) equalized odds requires that an AI has equal false positive and false negative rates across groups; (c) equal opportunity is satisfied when, among individuals who qualify for a positive outcome (e.g., a loan approval), the likelihood of selection is the same across demographic groups; (d) predictive parity is satisfied when predictions are equally accurate across groups by having the same positive predictive value, meaning that when the AI predicts a positive outcome, the probability of correctness is the same across groups; (e) individual fairness requires that similar individuals receive similar treatment based on relevant criteria regardless of their protected attributes; (f) counterfactual fairness checks that a decision wouldn't change if an individual's protected attribute (e.g., race or gender) were different, holding everything else constant. Beyond definitions of fairness in computer science: sometimes one considers that everyone receiving the same thing is fair (like when allocating offices in a law school), sometimes true randomness is fair (like in a lottery), and sometimes neither of these two is and what's fair depends on another notion of deservedness. In sum, the tradeoff between anything and fairness depends at least partly on what is fair for that particular decision.

A first problem for the tradeoff is that many fairness definitions incorporate accuracy, which means that fairness and accuracy can't be at odds for those definitions. Fairness in AI often refers to ensuring that algorithmic decisions don't unduly disadvantage certain groups. Several definitions of tie fairness to model accuracy measures—requiring that a model's performance (error rates, prediction

[22] Goce Ristanoski, Wei Liu and James Bailey, 'Discrimination Aware Classification for Imbalanced Datasets', *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (Association for Computing Machinery 2013) 1529–32 <https://doi.org/10.1145/2505515.2507836>.

[23] Bejamin Fish, Jeremy Kun and Lelkes, 'Fair Boosting: A Case Study' (2015) 1–4.

[24] Gabriel Goh and others, 'Satisfying Real-World Goals with Dataset Constraints', *Advances in Neural Information Processing Systems* (Curran Associates, Inc 2016) 1–9 <https://papers.nips.cc/paper_files/paper/2016/hash/dc4c44f624d600aa568390f1f1104aa0-Abstract.html>.

[25] Toshihiro Kamishima and others, 'Fairness-Aware Classifier with Prejudice Remover Regularizer' in Peter A Flach, Tijl De Bie and Nello Cristianini (eds), *Machine Learning and Knowledge Discovery in Databases* (Springer 2012) 36, 41–44.

[26] Muhammad Bilal Zafar and others, 'Fairness Constraints: Mechanisms for Fair Classification', *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (PMLR 2017) 962–69 <https://proceedings.mlr.press/v54/zafar17a.html>.

[27] Feldman and others (n 15) 259–68.; Ignacio Cofone, 'Algorithmic Discrimination Is an Information Problem' (2019) 70 Hastings Law Journal 1389, 1399–1404.

[28] Arvind Narayanan, " 21 fairness definitions and their politics." *Proc. conf. fairness accountability transp.,* Vol. 1170. 2018 (providing an overview).

reliability, etc.) is comparable across protected groups. When fairness is defined in terms of accuracy, the supposed tradeoff disappears because fairness itself demands accuracy.

Chiefly, equalized odds and predictive parity involve accuracy considerations. In both cases, fairness is measured in terms of how accurately the model performs for different groups, meaning that increasing fairness under these definitions doesn't reduce accuracy. Predictive parity requires that a model's precision is equal across groups. Among the individuals the model classifies as positive (true positives and false positives), the proportion who deserve the positive outcome (true positive) must be the same for every group. Predictive parity thus ties fairness to the accuracy of the positive predictions—it requires that no group is given positive labels that turn out wrong more often than another group. A "positive" prediction must carry the same likelihood of being correct for everyone: the accuracy rate of the algorithm's decisions must be consistent across groups. Equalized odds requires that an AI model produce equal true positive *and* false positive rates across different demographic groups—it's achieved when both a classifier's true positive rate and false positive rate are the same for each group. The first means that the model catches positives (e.g. qualified candidates) equally well for each group, and the latter means it makes mistaken choices at an equal rate for each group. Because equalized odds requires that accuracy-related measures are similar across subpopulations (it demands parity in the error rates), it relies on accuracy.

This is true for other definitions of fairness too. Equal opportunity, for example (a modification of equalized odds), is aimed at giving protected groups the same chance at positive outcomes when they actually qualify for them. It focuses only on the true positive rate being equal across groups odds: it enforces half of the equalized odds criteria (true positive rate parity but not false positive rate parity).[29] Equal opportunity ties fairness to the model's true positive rate (recall). By equalizing true positive rate, it requires that the model is equally *accurate* at identifying deserved beneficial classifications in each group. For example, if the prediction task is hiring, equal opportunity means the hiring model selects the same fraction of qualified applicants from every demographic group. Doing so addresses one aspect of accuracy (false negatives) by requiring that a disadvantaged group's qualified candidates aren't disproportionately overlooked.[30]

Recognizing that fairness metrics are subject to discretionary choices because they are imperfect approximations to the normative idea of fairness is helpful. But it often obscures that the same is true of choices over output variables and the normative idea of accuracy.[31] As a result, the decision over a fairness criterion can be wrongly framed as a tradeoff between an imperfect measure of a social good (a fairness metric) against a social good (accuracy). Really, the tradeoffs are among two imperfect metrics of social goods.[32] The outcomes of such tradeoffs are context-dependent, and the normatively desirable result hinges on the extent to which each of the metrics approximates better the desired social objective in each case.

b) Model Multiplicity

Conditional on any optimization choice, designers have a choice between models. Research on model multiplicity challenges the fairness-accuracy tradeoff in AI in one important way: by revealing that the deployed model is just one of many that could have been trained to achieve similar levels of output variable accuracy.[33] As Emily Black et al. explain, the assumptions "that a unique solution exists and that a fairness-accuracy tradeoff is inevitable—are descriptively inaccurate. Recent work in computer

[29] Hardt et al. (2016) Equality of Opportunity in Supervised Learning

[30] Unlike equalized odds, it doesn't constrain false positive rates, so it allows differences in how often the model mistakenly flags individuals who should have been classified negatively as positive cases.

[31] See Section 4.a

[32] Id.

[33] Charles Marx, Flavio Calmon and Berk Ustun, 'Predictive Multiplicity in Classification' in Hal Daumé III and Aarti Singh (eds), *Proceedings of the 37th International Conference on Machine Learning* (PMLR 2020).

science has established that there are almost always multiple possible models with equivalent accuracy for a given optimization problem."[34]

This work challenges the conventional argument that enhancing fairness inevitably sacrifices accuracy because it reveals that designers had choices.[35] Model multiplicity highlights that, in the process of training AI models, there isn't just one optimal model that balances accuracy and fairness—there are often many models that achieve similar levels of accuracy but differ significantly in how they treat different groups.[36] Among design choices were models that could have been trained with nearly the same accuracy but different distributions of false positives and false negatives.[37] And, while achieving the least discriminatory model while maintaining these metrics of accuracy may be difficult, finding a less discriminatory model rarely is.[38] This understanding suggests that fairness and accuracy aren't necessarily at odds—more equitable models can often be found without significant compromise on output variable performance.

These variations matter because the impact of false positives and negatives isn't evenly distributed across populations. Some groups, especially historically disadvantaged ones, often bear a disproportionate burden for an asymmetry between false positives and false negatives.[39] For instance, in a predictive policing model, one version might slightly overpredict crime in a particular community, leading to more false positives, while another version with similar accuracy might reduce this bias.[40] Model multiplicity implies that the model builders could have chosen a model that, while just as accurate, offers better distributional outcomes. It reframes the fairness-accuracy tradeoff as not an inherent limitation of AI but rather as a reflection of choices made during model selection. This suggests that the tradeoff between fairness and accuracy isn't as rigid as often portrayed, and more fair, less biased models are possible without significant loss of accuracy.[41] Related work on arbitrariness, also challenging the tradeoff, has indicated that reducing variance can achieve both output variable accuracy and near-fairness simultaneously without fairness interventions.[42]

4) Why the Tradeoff is a Myth

a) The Role of Output Variables

Bias in proxies is normally formally defined in terms of effects. For example, a pre-AI source defines proxy bias as one that "systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others."[43] A recent paper explains: "Despite differences in definitions,

[34] Black and others, 'Less Discriminatory Algorithms' (n 2) 4.

[35] Kim, Race-Aware Algorithms (n 6) at 1554; Alessandro Fabris and others, 'Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey' (arXiv, 8 April 2024) at 33, <http://arxiv.org/abs/2309.13933>.

[36] Black and others, 'Less Discriminatory Algorithms' (n 2) 30.

[37] Black and others, 'Less Discriminatory Algorithms' (n 2).

[38] ibid 4. ("Finding a specific equally accurate model—corresponding to a particular re-drawing of a model's decision boundary—is difficult through the typical model development process. At the same time, finding equally accurate models in general is easy, as they are common. Indeed, research has shown that models with equivalent accuracy which differ in other behavior (including disparate impact) can be easily discovered in practice, and occur naturally throughout the model development process [12, 27, 72]. As a result, while it is difficult to find the least discriminatory alternative model for any set of equally effective models, with some effort, a model that is less discriminatory than a baseline model can almost certainly be found in practice.").

[39] See Cofone and Strandburg, supra note X.

[40] Black and others, 'Less Discriminatory Algorithms' (n 2).

[41] ibid.

[42] A. Feder Cooper and others, 'Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification' (2024) 38th AAAI Conference on Artificial Intelligence 22004 at 22010. See also K Rodolfa and others, 'Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy' (2021) 3.10 Nature Machine Intelligence 896.

[43] Batya Friedman, Eric Brok, Susan King Roth, John Thomas, *Report: Minimizing Bias in Computer Systems*, CHI '95 Workshop

algorithmic bias generally refers to systematic unfairness or unequal treatment caused by algorithms, often disadvantaging or favoring groups based on immutable traits such as race, gender, or socioeconomic status."[44] But metrics of bias—what measures this *effect*—are usually implemented by measuring them against the algorithm's output variable. This overlooks that the output variable can be biased too.

Designing an AI algorithm requires defining a prediction objective (decision criterion), which in turn requires an algorithmic designer to distill private or social needs into metrics that can be processed by a model. In doing so, the chosen output variable is inevitably a proxy for a prediction objective. They are chosen because the ideal decision criterion—what society truly cares about—cannot be measured directly.[45] In the case of a hiring algorithm, decision-makers would (often) like to base hiring decisions on the likelihood that the job candidate will be a good employee. But data on whether someone will be a good employee at that job doesn't exist. So algorithmic designers (often) use the output variable "similarity with job candidates who have been hired" as a proxy for the likelihood that a person would get hired because that data may be the closest substitute.[46]

Output variables being simplifications of reality is true of automated decision-making, not just of employment. AI algorithms simply must rely on quantifiable metrics to make decisions. Given the intangibility of various social goals, algorithmic designers must select measurable stand-ins as output variables for them.[47] These output variables are intended to approximate various decision criteria, like when a company uses number of emails in a day to measure someone's productivity or number of likes to a product to measure customer satisfaction.

Social objectives are simply impossible to code as real-value functions.[48] The essence of these concepts, whether it's health, employability, or productivity, eludes direct measurement. As a result, models rely on tangible metrics as stand-ins, such as the number of steps walked for health or likes on a post for social approval. From a social point of view, the objective is always to maximize an abstract construct, such as "creditworthiness" and "teacher quality;"[49] it just happens that such construct must be operationalized by referring to a quantifiable metric instead. Using those metrics as stand-ins for social objectives may be inevitable and even acceptable. What it can't be is confused with those social objectives.

The very act of simplifying these criteria into quantifiable metrics introduces a layer of abstraction that can distort the algorithm's alignment with its intended purpose. Output variables capture a sliver of the intended concepts, introducing a disconnect with social values.[50] Many fairness issues, as well as other AI harms, stem from mismatches between unobservable theoretical constructs (e.g., "risk to society") and the output variables chosen to measure them.[51] The crux lies in the mismatch between these output variables and the decision criteria they aim to represent. Such mismatches are especially likely when training machine-learning-based algorithms because of the large volume of training data

[44] Conrad Borchers, "Toward Sufficient Statistical Power in Algorithmic Bias Assessment: A Test for ABROCA." *arXiv preprint arXiv:2501.04683* (2025). See also Batya Friedman and Helen Nissenbaum, Bias in computer systems, *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, 1996; Alexandra Jonker and Julie Rogers, What is algorithmic bias?, IBM, at https://www.ibm.com/think/topics/algorithmic-bias ("Algorithmic bias occurs when systematic errors in machine learning algorithms produce unfair or discriminatory outcomes.")

[45] Rachel L Thomas and David Uminsky, 'Reliance on Metrics Is a Fundamental Challenge for AI' (2022) 3 Patterns 1.

[46] Kim (n 6) 1551-1552.

[47] Shira Mitchell et al, 'Algorithmic Fairness: Choices, Assumptions, and Definitions' (2021) 8 Annual Review of Statistics and its Application 141 at 143.

[48] Black and others, 'Less Discriminatory Algorithms' (n 2) 8–9.

[49] Abigail Jacobs and Hannah Wallach, *Measurement and Fairness*, Proceedings of the ACM Conference on Fairness, Accountability, and Transparency 375 (2021) *At 380* <https://arxiv.org/pdf/1912.05511>.

[50] See Solon Barocas, Sophie Hood & Malte Ziewitz, *Governing Algorithms: A Provocation Piece*, at 4 ("new research might look at the way these companies frame the problems to which they then promise a solution. Rather than challenging the claims to efficacy, this work could reveal how problems are made suitable for an "algorithmic solution" in the first place").

[51] *Jacobs* at 375.

required. Datasets of the requisite size are only available for a limited selection of output variables that likely does not include the designer's ideal decision criterion.

This problem is related to the omitted payoff bias, where, by focusing on the (measurable) outcome being predicted, an algorithm can exclude other factors that decision-makers also care about,[52] but it exceeds it, because it pertains not *other* elements that decision-makers care about but the very element that the measurable outcome is sitting for. Consider the example of an algorithm designed to predict future managers within a company. If it uses past management appointments as a stand-in for leadership potential, it will perpetuate historical biases, overlooking qualified candidates who don't fit the traditional mold.[53] While the relevant output variable is the likelihood of being a good employee and one only has data for who was historically hired, that's an imperfect proxy for being a good employee. Data about hireability is biased by the fact that only some types of candidates were hired historically.[54] An individual might be less likely than others to be hired, resulting in a negative prediction, for example because of where they live or their social group even if, all things equal, they are as likely as others to perform well at the job.[55]

As a result, algorithm designers can face a different type of tradeoff: between the accuracy with which the machine learning model works *for a given output variable* and the reliability of that output variable *as a proxy for the ideal grounds* for decisions.[56] For example, as hires for any particular position are relatively rare, it's difficult to train a model to predict candidates for that position specifically because an appropriate sample of data for that output variable may be unavailable. Algorithm designers are forced to choose between less meaningful, but more numerous, hires for that company in general (or hires for that position across the industry) and the more meaningful, but less numerous, hires for that specific position at that company.[57]

b) An Example: Output Variable Bias in Recidivism Assessments

A recidivism algorithm is intended for assessing whether a defendant is likely to commit a violent crime if released on parole.[58] Risk-assessment algorithms combine information about an individual's prior arrests, prior employment, gang affiliation, and behavior in prison. Based on that information, these algorithms estimate the output variable "likelihood of rearrest if released on parole." In criminal justice, the fairness-accuracy tradeoff is presented as one between algorithmic fairness and releasing individuals who shouldn't be released.[59]

[52] See, e.g., Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S., 'Human decisions and machine predictions' (2017) 133 *Q. J. Econ.* 237.

[53] Pauline Kim, 'Auditing Algorithms for Discrimination' (2017) 166 University of Pennsylvania Law Review Online 189, 191.

[54] See generally Latanya Sweeney, 'Discrimination in Online Ad Delivery: Google Ads, Black Names and White Names, Racial Discrimination, and Click Advertising' (2013) 11 ACM Queue 10.

[55] See generally Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' (2019) 2 Columbia Business Law Review 494.

[56] John Nay and Katherine J Strandburg, 'Generalizability: Machine Learning and Humans-in-the-Loop' in Roland Vogl and Edward Elgar (eds), *Research Handbook on Big Data Law* (Edward Elgar Publishing 2020).

[57] Kim (n 6) 1544. ("The designers must make difficult choices each step of the way, involving tradeoffs, subjective judgments and the weighing of values. Each of these choices can be consequential in shaping the final model and the results it produces").

[58] Jessica M Eaglin, 'Constructing Recidivism Risk' (2017) 67 Emory Law Journal 59, 75.

[59] See, e.g., Alex Chohlas-Wood and others, 'Learning to Be Fair: A Consequentialist Approach to Equitable Decision-Making' <https://5harad.com/papers/learning-to-be-fair.pdf;>.; Sam Corbett-Davies and others, 'Algorithmic Decision Making and the Cost of Fairness', *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery 2017) 802–3.; Alex Chohlas-Wood and others, 'Designing Equitable Algorithms' (2023) 3 Nature Computational Science 601.

Whether a defendant will commit an offense if released is unknowable at the time of the decision.[60] The closest thing is to train a model to predict the output variable "likelihood that a defendant with particular demographics will be rearrested if released." The algorithm's output (that is, the likelihood of rearrest), of course, can under- or over-estimate the actual likelihood that a particular individual would be rearrested if released. More importantly, even if maximally accurate, the algorithm's output (the likelihood of rearrest) may under- or over-estimate the actual likelihood that a particular prisoner would *recidivate*.

This issue is far from hypothetical in recidivism prediction in particular because Black individuals are rearrested more often than white individuals. No matter how much high-quality data a recidivism model is trained on, if it's trained to predict the probability of being arrested, it will generate accurate arrest estimations that are biased recidivism judgements against Black individuals.[61] Training it on more arrest data will not help. For example, the Chicago Police Department's Strategic Subject List predicts the likelihood of a person being involved in gun violence in the future. However, because the list was allegedly used by the police as an informal suspect list for crimes involving gun violence, it was predictive not of involvement in future gun violence but of the probability of being arrested in the future.[62] Similarly, another risk assessment algorithm used at the federal level for probation was found to assign a higher average score of (post-conviction) risk assessment to Black individuals.[63] A study concluded that bias in this algorithm was unlikely because 66% of the racial difference was attributable to criminal history and, according to the study, because criminal history is not a proxy for race.[64] But criminal history does affect the relationship between race and future arrest.

There is abundant empirical evidence of arrests being racially biased.[65] According to one meta-analysis of quantitative research estimating the effects of race on the police decision to arrest, Black suspects are consistently more likely to be arrested than white suspects.[66] Some criminologists estimate that the likelihood of arrest is, on average, more than twice as high for Black suspects compared to white suspects,[67] while others put the gap in the probability of arrest between white and Black individuals between 9% and 22%.[68] Black individuals are arrested for violent crimes at a higher rate than white

[60] See, e.g., Julia Dressel and Hany Farid, 'The Accuracy, Fairness, and Limits of Predicting Recidivism' (2018) 4 Science Advances 1–5.

[61] Julia Angwin and others, 'Machine Bias' [2016] *ProPublica* <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

[62] Jessica Saunders, Priscillia Hunt and John S Hollywood, 'Predictions Put Into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot' (2017) 12 Journal of Experimental Criminology 363–64 <https://www.rand.org/pubs/external_publications/EP67204.html>.

[63] Jennifer L Skeem and Christopher T Lowenkamp, 'Risk, Race, and Recidivism: Predictive Bias and Disparate Impact' (2016) 54 Criminology 680, 685.

[64] ibid 700.

[65] Yu Du, 'Racial Bias Still Exists in Criminal Justice System? A Review of Recent Empirical Research' (2021) 37 Touro Law Review 85–91 <https://digitalcommons.tourolaw.edu/lawreview/vol37/iss1/7>.; Ngozi Okidegbe, 'Discredited Data' (2022) 107 Cornell Law Review 2007.

[66] Tammy Rinehart Kochel, David B Wilson and Stephen D Mastrofski, 'Effect of Suspect Race on Officers' Arrest Decisions' (2011) 49 Criminology 473, 490–91.

[67] Brendan Lantz and Marin R Wenger, 'The Co-Offender as Counterfactual: A Quasi-Experimental within-Partnership Approach to the Examination of the Relationship between Race and Arrest' (2020) 16 Journal of Experimental Criminology 183, 199.

[68] Stewart J D'Alessio and Lisa Stolzenberg, 'Race and the Probability of Arrest' (2003) 81 Social Forces 1381, 1388; Frank McIntyre and Shima Baradaran, 'Race, Prediction, and Pretrial Detention' (2013) 10 Journal of Empirical Legal Studies 741, 751; Emma Pierson and others, 'A Large-Scale Analysis of Racial Disparities in Police Stops across the United States' (2020) 4 Nature Human Behaviour 736.

individuals,[69] but get convicted at a lower rate.[70] When factors such as socioeconomic conditions, neighborhood crime rates, demographic turnover, and policing strategies are controlled for, Black individuals don't commit more violent crimes than white individuals, even though they are arrested at a higher rate.[71] The disparity cuts across types of arrests.[72]

Risk-assessment algorithms trained on arrest data, therefore, when not biased as a measure of rearrest. are biased as a measure of re-offense.[73] In using prediction of rearrest as a stand-in for re-offense, they pick up social biases from the historical data that distort the relationship between offending and being arrested.[74] In other words, the root of the problem is the human decision to use rearrest as a stand-in for re-offending.

An increment in false positives as a result of fairness techniques, the standard policy argument goes, is the fairness-accuracy tradeoff.[75] In criminal law, a standard policy approach to this issue, which has explicitly been the claim made about risk-assessment algorithms, posits that "society pays for fairness by having more dangerous individuals released."[76] However, even in a hypothetical scenario where these algorithms have maximal accuracy, people with a "high risk" output variable and people who will commit future crimes aren't perfectly overlapping groups: the datasets in most of these algorithms are racially biased because they use arrest data, which are racially biased concerning criminality.[77] This dynamic mirrors algorithms in employment, which may also produce skewed predictions because they reflect social bias by accurately reproducing real equity differences between groups.[78]

The data show that arrest is a biased measure of recidivism. The bias in arrest may even be larger than indicated by the data above. The preceding comparison of arrest and conviction rates demonstrates a racial bias in arrest assuming that convictions aren't racially biased, but that assumption is likely incorrect,[79] so the bias in the output variable "likelihood of being arrested" compared to the social construct "likelihood of committing a crime" is likely to be larger than reported bias in arrests.[80] The

[69] Federal Bureau of Investigations, 'Crime in the United States 2015' (*FBI*) <https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/tables/table-43/#overview>.

[70] Matthew R Durose, Donald Farole and Sean P Rosenmerkel, 'Felony Sentences in State Courts, 2006 - Statistical Tables' (2009) <https://www.bjs.gov/content/pub/pdf/fssc06st.pdf>. (National Judicial Reporting Program indicating that 39% of those convicted of violent crimes are Black and 58% are White).

[71] Gregory DeAngelo, R Kaj Gittings and Anita Alves Pena, 'Interracial Face-to-Face Crimes and the Socioeconomics of Neighborhoods: Evidence from Policing Records' (2018) 56 International Review of Law and Economics 1, 5–6. See also Daniel P Mears, Joshua C Cochran and Andrea M Lindsey, 'Offending and Racial and Ethnic Disparities in Criminal Justice: A Conceptual Framework for Guiding Theory and Research and Informing Policy' (2016) 32 Journal of Contemporary Criminal Justice 78, 80.; The Sentencing Project, 'Report to the United Nations on Racial Disparities in the U.S. Criminal Justice System' (2018).; Cydney Schleiden and others, 'Racial Disparities in Arrests: A Race Specific Model Explaining Arrest Rates Across Black and White Young Adults' (2020) 37 Child and Adolescent Social Work Journal 1.

[72] Du (n 51) 88–89.; Kochel, Wilson and Mastrofski (n 52).

[73] See, e.g., Corbett-Davies and others (n 43) 803, fig.2.; Jeff Larson and others, 'How We Analyzed the COMPAS Recidivism Algorithm' *ProPublica* <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

[74] Dressel and Farid (n 44).

[75] See, e.g., Marcus Tomalin and others, 'The Practical Ethics of Bias Reduction in Machine Translation: Why Domain Adaptation Is Better than Data Debiasing' (2021) 23 Ethics and Information Technology 419, 428.

[76] See, e.g., Chohlas-Wood and others, 'Learning to Be Fair: A Consequentialist Approach to Equitable Decision-Making' (n 43).; Sam Corbett-Davies and others, 'The Measure and Mismeasure of Fairness' (2023) 24 Journal of Machine Learning Research 1, 4.; Chohlas-Wood and others, 'Designing Equitable Algorithms' (n 43).

[77] Okidegbe (n 51).

[78] See Kim (n 36) 191.; Kim (n 6) 1548.

[79] See Du (n 51) 91–98.

[80] The magnitude of the bias would further increase if one considered that particular conducts are criminalized partly in their relationship to race and poverty, but the point on output variable bias would hold even ignoring such reality.

release of Black individuals who are more likely than others to be *arrested* in the future is therefore not necessarily a socially negative outcome, as it could be correcting for bias in the output variable.[81]

5) Implications

a) Why this Matters for Policy: AI Harm

This reframing of model accuracy matters for determining how to address AI bias. The choice of output variable isn't just a technical decision; it's a sociotechnical and ethical decision that carries implications for who benefits from AI and who is harmed by it. The gap between the output variable and the prediction objective that the output variable represents can lead to significant social consequences, such as reinforcing existing inequalities or prioritizing short-term efficiency at the expense of values like dignity or autonomy.

A common harm of the fairness-accuracy misunderstanding is individual discrimination. Often, there can be a gain in fairness without a loss in accuracy. In employment, for example, because "similarity" is used to estimate employability, an algorithm with lower predictive accuracy for its human-determined *output variable* (similarity) isn't necessarily any less accurate with regards to its prediction objective. Moreover, adding a fairness constraint to an algorithm that is known to be biased for the output variable may correct for the error between the chosen output variable and the target attribute (prediction objective), i.e., between the similarity with current employees and employability. Because false positives and false negatives are defined in terms of the output variable, and not in terms of the desired knowledge, an increment of false positives (an apparent loss in accuracy) is *not necessarily a bad thing*. Because the socially valuable metric is whether the job candidate will be a good employee (for which there are no data), an asymmetric false positive rate as to similarity with current employees may correct for the gender or racial bias in the output variable.[82]

Communities face harm too. When an output variable aligns poorly with its prediction objective, the consequences range from ineffectiveness to significant harm for communities.[83] AI not only replicates and perpetuates, but also amplifies, discrepancies between output variables and decision criteria. The iterative nature of machine learning, where algorithms continuously learn from data to improve their predictions, entrenches and magnifies the biases and limitations of the selected output variables. Without correction, automated systems spiral away from their intended social objectives, entrenching biased decision-making criteria. This failure to capture social determinants and magnifies historical biases reinforcing preexisting structural inequalities and marginalizing already disadvantaged groups.

More broadly, at the heart of output variable choices is the question of whose values and perspectives are encoded into the AI systems developed.[84] The selection of output variables, for that reason, isn't a value-neutral process; it reflects the priorities and biases of those who design and deploy these systems. As Pauline Kim explains, "because there is no single "correct" model for any given problem, there also is no "true" prediction for any given individual. The choices made in creating a machine learning model will affect the distribution of predicted outcomes, such that a particular person might score highly enough to receive a benefit under one model, but not under another, even before any group fairness considerations are considered."[85] Translating subjective social objectives into data points that an AI can optimize is a challenge at the core of AI development. And it introduces the choice about what to maximize. For example, when developing a healthcare diagnostic tool, a decision-

[81] Of course, it could under-correct or over-correct, depending on the size of the output variable bias and the size of the fairness adjustment,

[82] Anya Price and Daniel Schwarcz, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data' (2020) 105 Iowa Law Review 1257.

[83] Ngozi Okidegbe, 'The Democratizing Potential of Algorithms?' (2022) 53 Connecticut Law Review 739.

[84] Kim, 'Race-Aware Algorithms', at 1553.

[85] Kim (n 6) 1553.

making system could prioritize accuracy, speed, cost-effectiveness, or patient comfort. Each choice reflects a set of values being prioritized. This shows the subjectivity involved in choosing the output variable that will determine the goals and performance of a seemingly objective technology.

b) Why this Matters for Design

The inevitable reliance on output variables emphasizes the importance of transparency raised extensively in the literature—particularly but not only in the literature on algorithmic bias in employment.[86] Secrecy over output variables (beyond the opacity of many types of machine learning models) obscures the extent to which output variables drive decisions. This deliberate opacity masks potential misalignments between the output variable used and desired attributes, making it more difficult to identify and rectify mismatches and making the involvement of stakeholders more difficult.[87] The distance between social objectives and those functions is the source of many obstacles to interpretability.[88] The potential misalignment supports the argument that the output variables used should always be disclosed to the public, even overriding other considerations of trade secrecy and potential decision-subject gaming and even when there are legitimate concerns to maintain other parts of the automated decision-making secret.

The caution argued here regarding the accuracy and fairness of output variables supports prior arguments for assessing *consequential validity* in AI design, which involves examining an algorithm's real-world consequences and impacts on society.[89] Disparate and harmful impacts are apparent beyond the employment and criminal justice context because of how output variables are often determined by the training data, instead of the other way around. Fatalities caused by a self-driving cars illustrate the varied consequences of using imprecise proxies.[90] Similarly, a study of ImageNet (AI containing a database of web scraped images) revealed that seemingly neutral categories inherited racist and sexist connotations from WordNet (a database organizing and categorizing English words).[91] In the ImageNet example, relying on early 2000s search engines resulted in a dataset reflecting the racist and sexist biases prevalent online,[92] illustrating another context in which how output variables can pick up biases from by broad social phenomena. Mismatched output variables that skew the predictions of AI models, leading them to optimize for outcomes that diverge from social objectives, manifests in many other contexts, such as social media algorithms that prioritize engagement and as a consequence amplify sensationalist

[86] See, e.g., among others, Sánchez-Monedero, Javier, Lina Dencik, and Lilian Edwards 'What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems' (2020) *Proceedings of the 2020 conference on fairness, accountability, and transparency* 458; Andrew D. Selbst et al., *Deconstructing Design Decisions: Why Courts Must Interrogate Machine Learning and Other Technologies*, UCLA PUB. L. & LEGAL THEORY RSCH. PAPER NO. 23-22, at 60 ("[I]t is imperative that courts start to deconstruct design choices in ML . . . technologists need to be far more transparent about the nature of the choices they make when designing new technology. They must create detailed documentation of the design choices made and the rationales for them."); Stephen Casper et al., *Black-Box Access is Insufficient for Rigorous AI Audits*, PROC. OF 2024 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 2254 (2024); Karl Werder et al., *Establishing Data Provenance for Responsible Artificial Intelligence Systems*, 13 ACM TRANSACTIONS ON MGMT. INFO. SYS. 22 (2022); Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC'Y 1, at 9 (2016) ("Finding ways to reveal something of the internal logic of an algorithm can address concerns about lack of 'fairness' and discriminatory effects, sometimes with reassuring evidence of the algorithm's objectivity."); Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 *INT'L DATA PRIV. L.* 76, 91–92 (2017) (highlighting the value of transparency in identifying and addressing algorithmic errors to improve societal outcomes).

[87] *See generally* Abebe (n 54)

[88] Zachary C Lipton, 'The Mythos of Model Interpretability' (2018) 61 Communications of the ACM 36, 2.

[89] Abigail Z. Jacobs & Hanna Wallach, *Measurement and Fairness*, PROCEEDINGS OF THE ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 375, 375 (2021) <https://arxiv.org/pdf/1912.05511>.

[90] Alex Hanna et al, *Lines of Sight*, 12 LOGIC(S) (December 20, 2020) https://logicmag.io/commons/lines-of-sight/?s=03.

[91] Alex Hanna et al, *Lines of Sight*, 12 LOGIC(S) (December 20, 2020) https://logicmag.io/commons/lines-of-sight/?s=03.

[92] Id.

or divisive content. Such outcomes don't misalign with the intended social goals at random; they tend to perpetuate existing biases and inequalities in their deviation.

Lastly, this matters for participatory design. The need for large volumes of data to train algorithms often means that output variables are chosen based on availability rather than suitability. Because gaps between output variables and its prediction objectives affects decision-subjects and communities as much as decision-makers, the discussion on which criterion serves as an acceptable basis for a particular type of decision should involve those decision-subjects and communities as well as decision-makers.[93] Addressing the mismatch between output variables and the decision criteria in machine learning requires critical evaluation not just of the output variable's statistical usefulness but also their social usefulness. That requires asking questions that go beyond data availability and the quantifiability of each metric.

c) Why this Matters for Antidiscrimination Law: Statistical Discrimination

Under U.S. law, algorithmic discrimination is mostly a disparate impact problem,[94] although this isn't always the case.[95] Accuracy-independent fairness constraints are means to prevent statistical discrimination as defined by the disparate impact doctrine because disparate impact discrimination is a problem of adversely affecting protected populations without a classification bias. Consequently, to address these biases, entities that must comply with disparate impact discrimination under employment, financial, and housing law should be required to apply fairness constraints. Courts and administrative bodies, even under disparate impact principles, have stated that less discriminatory alternatives must be equally or comparably "effective."[96] Whether the fairness-accuracy tradeoff is false, under that logic, is determinative of whether alternatives are equally effective.The fact that there's often no accuracy loss when imposing fairness constraints shows that there is no business justification not to impose them.

Disparate impact doctrine prevents statistical discrimination in practices that, while appearing neutral, disproportionately affect certain groups. In employment, housing, and lending, if a policy leads to statistical disparities among protected classes, it can be deemed discriminatory unless a valid business necessity exists. In lending, for instance, a credit algorithm that rejects more loans for a particular group it might be considered discriminatory even without explicit bias. Unjustified discrimination can be recognized and rectified even if it was done unintentionally.

For the fairness-accuracy tradeoff to be possible, a model that makes accurate predictions must lead to disparate impact discrimination. One common form of algorithmic discrimination is the result of using biased samples.[97] But accuracy claims require that a model is trained with representative data. To understand why a model trained with representative data may not comply with algorithmic fairness metrics one can turn to the idea of statistical discrimination.[98]

Another form of inequity that is relevant, statistical discrimination, involves determining why, in a model trained with representative data, two individuals with the same productivity signal from different groups might be treated differently.[99] This form of statistical discrimination is sometimes an issue of

[93] Rediret Abebe and others, 'Roles for computing in social change' (2020) *Proceedings of the 2020 conference on fairness, accountability, and transparency* 252 at 255.

[94] Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671, 701–12.; Kim (n 6) 1557.

[95] Stephanie Bornstein, 'Antidiscriminatory Algorithms' (2018) 70 Alabama Law Review 519, at 564; Hilde Weerts and others, 'Unlawful Proxy Discrimination: A Framework for Challenging Inherently Discriminatory Algorithms' (2024) *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* 1850.

[96] Emily Black and others, 'The Legal Duty to Search for Less Discriminatory Algorithms' (n 5) at 8.

[97] Kim, supra note X.

[98] Edmund S Phelps, 'The Statistical Theory of Racism and Sexism' (1972) 62 The American Economic Review 659, 661.; Kenneth J Arrow, 'The Theory of Discrimination' in Orley Ashenfelter and Albert Reeds (eds), *Discrimination in Labor Markets* (Princeton University Press 1973).

[99] See Shelly Lundberg and Richard Startz, 'On the Persistence of Racial Inequality' (1998) 16 Journal of Labor Economics 292, 292–95.

differential observability. Differential observability takes place when skill distributions in two groups are the same, but the signals for individuals in group A are noisier (i.e., weaker or less informative) than those of individuals in group B. This will lead an algorithm to consider the expected productivity of a worker from group A, with its noisier signal, to be closer to the population average than the expected productivity of a worker from group B with an equivalent (although less noisy) signal because the algorithm will allocate different weights to signals for members of each group. This difference will lead highly qualified individuals from group A to receive a lower salary than their equivalents from group B.[100] Other times, this form of statistical discrimination is an issue of stereotyping. Stereotyping is a type of differential treatment that takes place when productivity signals are equally noisy (i.e., equally strong or informative) for individual members of both groups, but the first group has a lower average human capital investment than the second. Because the model will consider both group membership and the individual signal to be informative of each individual's expected productivity, it will consider an employee from the first group to have lower *expected* productivity than an employee from the second group who has the same signal. Therefore, workers from the first group may receive a lower salary under the same signal, or they will be offered fewer jobs, which constitutes disparate impact.

Things would be different if the tradeoff existed across the board. Imagine one held the (mistaken) belief that the algorithm discussed above was not disadvantaging members of a protected category by categorizing more members of that category as of low productivity and, therefore, one were to believe that training the algorithm with fairness-adjusted data would lead to lower predictive accuracy—the policy argument discussed above. In that case, whether one would consider an algorithm as discriminatory would depend on whether there are duties of reasonable accommodation, where the accommodation would be to incur a loss in accuracy to avoid discrimination.[101] Whether a duty to accommodate exists, or whether the law cannot impose a reduction in accuracy, would arguably hinge upon which antidiscrimination principle is used—disparate treatment or disparate impact. In some circumstances, the law under disparate impact doctrine requires decision-makers (in this case, an employer), to provide reasonable alternatives to perform a task even at a cost to the decision-maker.[102] So, while when the antidiscrimination principle used is disparate treatment, imposing costs on the decision-maker when facially neutral rules are employed is unwarranted, when the principle used is disparate impact, one should expect the law to introduce some cost, such as reduced overall accuracy, in a similar way that disparate impact antidiscrimination law does for other decisions.

6) The Tradeoff Revisited

a) When the Tradeoff is Possible

A fairness-accuracy tradeoff can exist under specific conditions. For there to be a tradeoff, three requirements must be met: the definition of fairness used must not explicitly incorporate accuracy, the output variable must be a reliable measure of what the model attempts to predict, and the output variable must not be biased. If any of these conditions aren't satisfied, then fairness can potentially be improved without sacrificing accuracy. Only when all three conditions hold simultaneously increasing fairness necessarily reduces predictive performance with regards to the prediction objective.

The first condition is that the fairness definition in question must not require accuracy. Some fairness metrics, such as equalized odds or predictive parity, explicitly incorporate accuracy-related measures by requiring that error rates or predictive values be equal across groups.[103] In such cases, improving fairness doesn't reduce accuracy because fairness itself is tied to ensuring the model performs well across demographic groups. However, some fairness measures, such as demographic parity and

[100] Hanming Fang and Andrea Moro, 'Theories of Statistical Discrimination and Affirmative Action: A Survey' in Jess Benhabib, Alberto Bisin and Matthew O Jackson (eds), *Handbook of Social Economics*, vol 1A (2011) 137–40.

[101] Kim (n 6).

[102] Christine Jolls, 'Antidiscrimination and Accomodation' (2001) 115 Harvard Law Review 642, 697–99.

[103] See Section 3.a.

equal selection rate, don't incorporate accuracy considerations. These do impose constraints that can force a model to make predictions in ways that can round counter to making correct classifications. If fairness is defined in terms of equalizing outcomes without considering correctness, then achieving fairness may require reassigning positive or negative predictions in ways that lower the model's predictive reliability. A fairness-accuracy tradeoff can only exist when the chosen fairness definition permits trade-offs by overriding accuracy.

The second condition is that the model's output variable must be closely tied to the ground truth it seeks to measure. If the relationship between a model's output variable and the underlying prediction objective is noisy, then modifying the model to improve fairness may not meaningfully reduce accuracy, because accuracy in a meaningful sense was low to begin with (even if the measured model accuracy with regards to its output variable was high). When the output variable closely tracks the prediction objective, altering its decisions for fairness reasons in the sense of the last paragraph can decrease predictive performance. For example, if a credit scoring model is based on reliable financial indicators that strongly predict loan repayment, adjusting it to ensure equal loan approval rates across demographic groups can lead to increased defaults. Similarly, in medical diagnostics, if a model predicts disease risk based on biomarkers with high predictive validity, imposing fairness constraints that require equal proportions of positive classifications across demographic groups can reduce its ability to correctly identify sick patients. The closer the connection between the model's output variable and the ground truth the model attempts to capture, the greater the risk that accuracy-independent fairness constraints will distort this connection and result a loss of predictive performance with regards to the prediction objective.

The third condition is that the output variable itself must not be biased. If the data with which the model's accuracy is measured is already distorted by societal bias—such as biased past hiring decisions for a similarity output variable or biased arrest data for a probability of re-arrest output variable—then enforcing fairness constraints (even accuracy-independent ones) may actually correct these distortions rather than introduce new errors even if the output variable doesn't have random noise embedded. If the goal is to predict an outcome that has had predictions with a verifiable skew in the past, prioritizing fairness can improve the model's real-world reliability by aligning its predictions with the underlying ground truth rather than with biased historical data. It is only if the output variable is unbiased and differences exist across ground with regards to the ground truth, not merely the output variable, that forcing the model to equalize outcomes across groups means distorting accurate predictions. In a context where the recorded data to measure output variable accuracy is already fair, and the output variable is not noisy, modifying the model to achieve accuracy-independent fairness metrics reduces predictive performance, because it forces the model to ignore real statistical differences.

In sum, if a fairness metric incorporates accuracy, then no tradeoff exists because fairness is defined in terms of maintaining predictive performance. If the model's predictions are weakly related to the actual outcome, then modifying them for fairness reasons doesn't necessarily reduce accuracy. If the output variable itself incorporates bias, then improving fairness may improve the model's validity. Only when fairness is defined in a way that disregards accuracy, when the model has a strong predictive relationship with the true outcome, and when the output variable is unbiased does increasing fairness mean reducing predictive performance. Any argument for an inevitable fairness-accuracy tradeoff in a specific setting should show that all three conditions hold; otherwise, improving fairness might leave accuracy unchanged, or even improve it, rather than worsen it.

b) An Example: Dermatology Imaging

An example where the fairness-accuracy tradeoff might happen, if these three conditions are fulfilled, is AI models for skin cancer detection. These models, trained on datasets of labeled medical images, predict whether a skin lesion is malignant (e.g., melanoma) or benign. Their goal is to maximize diagnostic accuracy so that patients at high risk receive timely treatment while minimizing unnecessary biopsies for those with benign lesions.

The first necessary condition is that the fairness constraint must not incorporate accuracy. A commonly proposed fairness metric in AI is demographic parity, which would require the model to predict skin cancer at equal rates across different demographic groups, such as individuals with lighter versus darker skin tones. Unlike other fairness measures such as equalized odds, which would check that false positive and false negative rates are equal across groups, demographic parity wouldn't consider the correctness of predictions. Instead, it would make the model classify the same proportion of cases as malignant across demographic groups, even if some groups statistically have different skin cancer prevalence rates. Enforcing demographic parity in a skin cancer detection model would mean that if lighter-skinned patients have a higher base rate of melanoma (not only in the training data but also out in the world), the model would inflate its malignancy predictions for darker-skinned patients to match the same positive classification rate. Or it may also lower its positive classification rate for lighter-skinned patients to equalize the overall distribution of predictions. Since demographic parity doesn't ensure that predictions align with the disease prevalence, it could introduce diagnostic errors—leading to either an increase in false positives (unnecessary biopsies) or false negatives (missed cancer diagnoses).

The second necessary condition is that the model's output variable—skin cancer presence—is closely tied to what it attempts to measure. These imaging models are trained on enormous datasets of high-resolution medical images labeled by experts. The relationship between visual features (such as lesion asymmetry, border irregularities, and color variation) and cancer presence is well-established in medical research, making these models a helpful tool for skin cancer detection. Some of them can match or exceed the accuracy of an eye exam by a dermatologist in identifying melanoma. These models' predictions aren't arbitrary—they are grounded in biologically meaningful features that correlate with actual disease. Because these model's predictions are strongly tied to objective clinical markers of disease, overriding them for fairness reasons can harm diagnostic accuracy. If the opposite were true, and a model were trained on unreliable data that would make its predictions with, say, 50% accuracy, introducing demographic parity might not be a problem, but this is not the case for most imaging models.

The third condition is that the output variable (whether a lesion is cancerous) isn't biased. Unlike hiring and criminal risk assessment, where data over historical human decisions introduce bias into the output variable itself (hireability or high likelihood of arrest), cancer presence is an objective biological fact. Confirmed diagnoses serve as the ground truth for labels in the training data, so the output variable "similarity between this image and other images labeled as cancerous" is unbiased. Because melanoma occurrence varies naturally across different skin types due to genetic and environmental factors, real differences exist between demographic groups. Although disparities in healthcare access and physician bias in clinical diagnoses do exist, the final diagnosis from a biopsy is reliable as it's determined by histopathological examination under a microscope. Enforcing demographic parity constraints that override differences introduces distortions rather than correcting bias only if the AI model is trained on accurate, unbiased diagnostic labels and reflects real differences in its predictions; but if the model were trained on biased data, such as initial physician assessments or mostly images from light skin tones, the constraints could potentially improve accuracy by compensating for those biases.

This case therefore could satisfy all three conditions that make the fairness-accuracy tradeoff real. The output variable (similarity with images diagnosed as having melanoma) is closely tied to what the model attempts to measure (melanoma presence); and the output variable isn't biased (it's based on medically-determined pathology), so if a fairness constraint that doesn't account for accuracy is imposed, it would degrade this model's predictive performance. That notion of fairness would be at a tradeoff with a meaningful notion of accuracy.

This example is uncommon. How specific the tradeoff circumstances need to be shows that the fairness-accuracy tradeoff isn't a feature of all AI models but rather a consequence of specific fairness definitions, predictive relationships, and unbiased ground-truth variables. If any of these conditions were absent—if fairness incorporated accuracy, if skin cancer were only weakly correlated with image-based predictors, or if the cancer labels themselves were biased—then the tradeoff might not exist, and accuracy-independent fairness interventions might not affect or improve performance.

7) Conclusion

The claim that AI fairness and accuracy are necessarily at a tradeoff is a myth. And the corollary policy concern often invoked in law about the accuracy cost of fairness measures is therefore overblown. The false perception of a blanket tradeoff between accuracy and fairness oversimplifies conclusions from algorithmic fairness and overlooks ethical aspects of how algorithms are deployed. Output variables are measurable constructs that sit in for the unmesurable ones that decision-makers ultimately aim to optimize. So AI decision-making requires an implicit or explicit evaluation of the alignment of output variables with private and social objectives, as they do of fairness criteria.

The first side of the equation is fairness: The tradeoff can only take place when the fairness definition requires overriding statistical relationships that are strongly predictive and unbiased. Demographic parity and treatment equality fall into this category because they force a specific outcome distribution rather than ensuring fairness in error distribution or probability calibration. The tradeoff between fairness and accuracy depends on which fairness definitions and output variables are chosen. Only if fairness is defined in a way that can conflict with an accuracy metric (e.g., demographic parity), enforcing fairness reduces accuracy. If fairness is defined in terms of accuracy measures (e.g., equalized odds vs. false positive/negative rates), then the tradeoff is nonexistent. As a result, fairness and accuracy aren't always at a tradeoff as a matter of logical necessity.

The choice of output variable is the other side of the equation—regarding accuracy. The broad term "algorithmic proxy" should refer to two meanings: (a) the labels of a machine learning algorithm, as commonly used, and (b) the output variable of a machine learning algorithm. In both cases the proxy is, by definition, never the actual target. Understanding the gap between algorithmic decisions and social values requires a closer connection between legal scholarship and computer science research on the limitations of fairness metrics. But, more than that, closing that gap for each decision requires a broader sociotechnical approach so that the chosen output variables align as closely as possible with social values. Moreover, this gap provides an argument for transparency in algorithm design to allow for scrutiny of the values each model intendedly or unintendedly promotes. As a result, because metrics of accuracy can be chosen strategically to ignore definitions of fairness, a mere conflict shouldn't be an excuse to ignore AI fairness.

Output variable mismatch in AI illustrates a broader issue. As we increasingly delegate decision-making to algorithms, the need to critically examine and refine the prediction objectives that guide these decisions is key. This effort requires reconciling AI applications with the social objectives they purport to serve. Understanding this context-dependent relationship is crucial for making choices about fairness interventions, fairness metrics, and accuracy assessments.