Accuracy Isn't Admissibility

Proxy Discrimination and the Limits of the Engineer's Defence

Name: Matthew Hinz

Student Number: 21228453

Phil 673

Fairness and Anti-Discrimination in AI Algorithms

March, 2026

Section 1: Introduction

Predictive machine-learning systems increasingly sort people in high-stakes screening settings. Call this the engineer's defence: predictive systems learn from an already unequal world, and because engineers do not control the institutions that produced that inequality, their task is narrower, optimise predictive accuracy rather than try to repair society upstream. That position is worth taking seriously because automated decision tools are often defended as more accurate or less biased than ordinary human judgement (Ajunwa, 2020, pp. 1671, 1673–1675; Kelly-Lyth, 2023, pp. 152–153).

That framing becomes unstable once we look at proxy discrimination. A system may exclude race or sex as formal inputs and still recover much the same social information through zip code, school attended, spending habits, or other facially neutral variables. As Prince and Schwarcz argue, proxy discrimination is not just any disparate impact but a case in which a facially neutral feature is useful at least in part because it tracks protected-class information or the social meaning attached to it; with contemporary AI, blocking the most obvious proxies does not solve the problem so much as drive the system toward less intuitive ones (2020, pp. 1260–1264). Johnson makes the same point in philosophical terms: biases that operate through proxies cannot be eliminated simply by filtering out explicit references to protected traits, because the decision procedure can substitute other correlated features instead (2021, pp. 9941–9943).

The issue is not simply whether a variable improves prediction, but whether predictive usefulness by itself makes that variable a permissible basis for allocating opportunities. The engineer's defence assumes that it does. Yet algorithmic systems do not merely read the world; they are built around contestable choices about objectives, outcome measures, predictors, and trade-offs. As Kleinberg et al. argue, screening systems are structured around explicit goals, and with adequate record-keeping they can make visible the decisions and trade-offs built into their

construction (2019, pp. 114–115). Even 'accuracy' is not neutral when the target itself is shaped by prior inequality (Kelly-Lyth, 2023, pp. 155–156).

This paper argues that proxy discrimination shows why the engineer's defence fails, because it silently assumes that predictive validity is enough to make a variable admissible in high-stakes selection, and that assumption does not hold. The later sections show why this matters especially in bottlenecked screening domains and what follows from it institutionally. Section 2 defines the engineer's defence in its strongest form. Section 3 explains proxy discrimination and why removing protected variables does not solve it. Section 4 argues that proxy discrimination defeats the defence by exposing its hidden bridge principle. Section 5 shows why the critique is especially sharp in bottlenecked screening domains such as hiring and admissions. Section 6 then asks what would count as a justifiable response.

Section 2: The Engineer's Defence

What I am calling the engineer's defence is not a settled term of art but a recurring justificatory pattern in contemporary debate. Automated systems are introduced as a way to reduce the role of flawed human judgement; they are sold as more consistent, more objective, or less biased than ordinary discretion; and when bias reappears, responsibility is often pushed backward into the world from which the data were drawn. Ajunwa describes this rhetoric clearly: automated decision-making is often framed as an anti-bias intervention because removing humans from the process is supposed to remove their bias as well, even though in practice such systems can reproduce and amplify inequality (2020, pp. 1673–1675). Kelly-Lyth makes the same point in the employment context, noting that arguments about algorithmic management routinely proceed through claims that algorithms are 'more accurate' or 'less biased' than human decision-makers, and the engineer's defence gives this rhetoric its most disciplined form (2023, pp. 152–153).

This defence has three steps. First, predictive systems learn from historical data, and historical data often encode the residue of discrimination. Second, engineers do not control the upstream institutions that produced those patterns: schools, labour markets, housing, policing, family wealth, and the rest. Third, because engineers cannot repair society upstream, their proper role is narrower: they should build the most accurate predictor they can and avoid imposing further distortions of their own. On this view, deliberate departure from predictive fit looks presumptuous because it asks engineers to substitute contested moral judgements for whatever signal the world has made available, and that position begins from two truths: the data are not innocent, and the engineer did not create the world from which the data came.

At its most plausible, the defence appears as a claim about role and epistemic humility. It rejects the fantasy that a model builder can simply wash injustice out of a dataset and deliver a purified social world in code, while also explaining why accuracy has such grip in high-stakes settings. If an employer must choose among applicants, or a university among candidates, 'accuracy' appears to offer a disciplined way of tying selection to some declared institutional goal rather than to whim or prejudice. Kelly-Lyth shows why this line of thought is so attractive in practice: once human decision-making is taken to be opaque, inconsistent, and biased, the algorithm is cast as the cleaner alternative (Kelly-Lyth, 2023, pp. 153–155). In that respect, the defence is not a straw man but an expression of a serious worry about arbitrariness.

Even so, the defence is incomplete even before any deeper normative objection is introduced, because algorithms do not merely mirror the world but are built to solve a specified problem under a specified objective, using selected inputs, target variables, and deployment rules. Kleinberg et al. make this especially clear: their point is not just that algorithms can be audited after the fact, but that the process of construction itself contains identifiable choices. They therefore propose record-keeping requirements that would preserve "valuable transparency" about "the decisions and choices made in building algorithms" and the trade-offs among values embedded in those

choices (2019, p. 114). The same article defines screening decisions as cases in which institutions choose from a pool of candidates for a particular goal, for example admitting students on the basis of academic potential and emphasises that the predictive system is organised around that prior specification of the task (Kleinberg et al., 2019, p. 115). So even before one can ask whether a predictor is fair, one must ask what it has been built to predict, why that outcome has been selected, and what kinds of evidence are being allowed to count.

O'Neil states the same point in a more directly: building a model requires deciding what is important enough to include, what can be ignored, and what will count as success. A model's blind spots therefore do not sit outside the exercise of judgement; they are one of the places that judgement enters it. In that sense, the instruction to 'just optimise accuracy' is already normatively freighted, because it presupposes a contestable view of which variables may stand in for the world and which institutional objective is worth optimising for (O'Neil, 2016, p. 6).

Martin adds the accountability point: once algorithms structure decisions and distribute roles within them, firms remain responsible for the ethical consequences of that design: they create moral consequences and structure who does what within a decision process (2019, pp. 835–837). That is enough, for present purposes, to block the thought that the engineer merely receives a contaminated world and transcribes it as faithfully as possible. The engineer still helps decide how the world is to be represented, which patterns are to be treated as informative, and what institutional aim the system will serve, and while none of that yet shows that the defence fails, it does show that 'just optimise accuracy' is not a purely technical instruction but one that already rests on contestable choices about objectives, outcome proxies, and admissible predictors.

Section 3: Proxy Discrimination

Proxy discrimination is narrower than disparate impact. Prince and Schwarcz define it as a case in which a facially neutral practice disproportionately harms a protected class and is useful

at least in part because it does that work, such that the feature's predictive value is bound up with the protected information it transmits or reconstructs rather than with a merely accidental correlation (2020, pp. 1260–1261). A bottom-line disparity may be accidental, whereas in proxy discrimination the disparate effect is part of what makes the feature useful.

That distinction matters because not every case of actuarial fit is proxy discrimination, and not every case of proxy discrimination depends on animus. Historically, proxy discrimination was easiest to see in intentional forms such as redlining, where the proxy was chosen because it covertly tracked the forbidden trait. But the same structure can arise without hostile intent whenever a protected characteristic is predictive of some facially neutral goal and the system latches onto a stand-in for it (Prince & Schwarcz, 2020, pp. 1261–1263). The category therefore includes what they call unintentional but rational proxy discrimination, rational not in the moral sense but from the standpoint of prediction, and that is the form that matters most here because the engineer's defence is not built to defend overt bigotry but the use of predictors that seem justified because they improve fit.

Prince and Schwarcz use as an example a life insurer whose pricing model charges more to applicants who belong to a Facebook group focused on increasing access to BRCA testing for African Americans. Membership in that group may correlate with genetic predisposition to certain cancers, and so it may improve prediction of future claims. In one sense, then, the feature is actuarially useful, but that does not settle the question. If the reason the feature helps is that it routes through genetic information or its social residue, then the model is not merely differentiating risk in the ordinary sense. It is using a facially neutral variable as a stand-in for protected or legally suspect information (Prince & Schwarcz, 2020, pp. 1261–1262). That answers the actuarial-fairness worry, because fit and permissibility come apart. A feature can improve pricing while still being objectionable because of the kind of information it conveys and the kind of classification it reconstructs.

The example also shows that proxy discrimination is not identical to any disparate impact produced by a neutral-seeming rule. Prince and Schwarcz note that the insurer in their example would likely be proxy discriminating with respect to genetic information, but not necessarily with respect to race. The Facebook group's race-specific character might create a racial disparate impact, but that alone would not show that race is doing the relevant explanatory work; if the feature helps because it tracks genetic predisposition rather than race as such, the racial impact is downstream and fortuitous relative to the model's predictive logic (Prince & Schwarcz, 2020, pp. 1261–1262). This boundary matters because proxy discrimination does not mean any case in which protected groups are worse off, but one in which the usefulness of a feature derives from the protected information it carries, reconstructs, or substitutes for.

That boundary also shows why simply removing protected variables does not solve the problem. As Tschantz emphasises, proxy discrimination is not one simple thing: some accounts focus on whether a feature has the capacity to recreate a protected trait, while others focus on whether the model actually uses that capacity in producing an outcome (Tschantz, 2022, pp. 1–4). For present purposes, the point is just that a feature does not become a relevant proxy because it is abstractly correlated with a protected trait; it must also matter in the model, and once it does, blindness is unstable.

Johnson describes this as the Proxy Problem. Proxy attributes are seemingly innocuous features that correlate with socially sensitive ones and attempts to suppress them create a dilemma: the model can often substitute another correlated feature, while more aggressive suppression risks reducing judgement accuracy (2021, pp. 9941–9943). Prince and Schwarcz push the point further: when directly predictive but legally suspect information is unavailable, AI systems are structurally driven to seek out substitute correlations, and blocking the most intuitive proxies often just forces the model toward less intuitive ones (2020, pp. 1263–1264). Proxy

discrimination therefore shows why formal exclusion is unstable: a model may still purchase accuracy through features that reconstruct protected status indirectly.

Section 4: Why Proxy Discrimination Defeats the Engineer's Defence

The engineer's defence fails at a single step: it begins with an epistemic claim that some variable or combination of variables improves prediction and ends with a normative entitlement to use that variable as a reason for allocating opportunities. That suppressed bridge principle says, plainly, that sufficiently strong predictive validity confers normative admissibility in high-stakes selection. Proxy discrimination matters because it forces that principle into view: once it is clear that prediction can improve by routing through protected status or its social residue, the defence can no longer present itself as a merely technical posture of humility but must defend a substantive claim about what kinds of reasons institutions may act on.

Hellman gives the cleanest diagnosis of where the move goes wrong: the most prominent fairness measures do not all answer the same kind of question, since some concern what one ought to believe about a person given a score while others concern how persons are treated. Equal predictive accuracy, on her account, ensures that a score means the same thing across groups, and so bears on belief. But fairness ordinarily concerns action, not belief (Hellman, 2020, p. 812). That distinction cuts directly against the engineer's defence, because a model may identify statistically informative features and even generate well-calibrated beliefs without it following that those same features may permissibly be used to hire, reject, rank, exclude, or otherwise distribute opportunities; predictive success answers one question, permissible treatment another.

That gap becomes sharper in hiring and admissions, where prediction is often redescribed as merit. Fishkin's formulation is exact: "It is not defensible to define 'merit' circularly, as performance on whatever tests we happen to have at hand" (2014, p. 57). The point is not that prediction is irrelevant, since institutions obviously need ways of selecting among candidates, but

that predictive fit cannot by itself settle what the institution is entitled to treat as a merit-relevant reason. If a university predicts future academic performance partly through school pedigree, neighborhood, recommendation style, or other variables saturated by prior inequality, the fact that those variables help prediction does not show that they track merit in the relevant sense. It shows only that they help sort within an already structured social world, and the same is true in employment, where a variable can improve prediction because it tracks advantages that were themselves unevenly distributed and normatively contested.

Anderson helps state what is at stake when those variables are used anyway, because on her account the point of equality is not to neutralise luck in the abstract but to end oppression and create a community in which people stand in relations of equality to others (1999, pp. 315–317). In that frame, the problem with proxy-laden prediction is not exhausted by error rates or even by disparate impact. The deeper question is whether the grounds of prediction are acceptable reasons for distributing standing, opportunity, and social position. Hiring and admissions are not simply moments of information processing but institutional sites where some people are admitted to paths of development and others are turned away. To treat a proxy for protected status as a reason within those decisions is to treat stratified social facts as admissible grounds for allocating future advantage.

Once that is seen, the engineer's defence loses its normative cover, because it can no longer say that the model merely found signal and engineers merely followed the data: the question is not whether the data contain signal, but whether the signal is of the right kind to authorize action. Proxy discrimination shows that some predictive success is purchased through informational routes that law and morality already have reason to regard with suspicion. Hellman explains why belief does not settle treatment. Fishkin explains why prediction does not settle merit. Anderson explains why the stakes are relational and political, not merely technical. Together, they block the

bridge principle on which the engineer's defence depends. This does not yet settle the institutional response, but it is enough to rule out a simple appeal to accuracy.

Section 5: Screening, Bottlenecks, and the Backwards-Looking Problem

The force of the preceding critique becomes clearest in screening domains such as hiring and admissions. Kleinberg et al. define screening decisions as cases in which an institution chooses from a pool of candidates to achieve some stated goal, as when students are admitted on the basis of academic potential or applicants are selected on the basis of predicted future performance. That framing clarifies that because these systems are not simply estimating facts about the world but helping decide who gets access to scarce openings that matter for what comes next (Kleinberg et al., 2019, p. 115). A university seat, an internship, a first job, a promotion: each is a site where institutions sort persons into paths that will shape later options. Once predictive models enter those settings, the question is no longer whether the model captures information well but whether the information on which it relies is an acceptable basis for allocating access to the next stage of opportunity.

This is where the backwards-looking problem bites, because as Mayson argues in a different domain, the deep problem is not merely dirty data or a flawed algorithm but the nature of prediction itself: all prediction looks to the past to make guesses about the future, and in a stratified world that means projecting the inequalities of the past into the future (2019, p. 2218). In hiring and admissions, the inputs to prediction are already saturated with earlier allocation decisions. School attended, grades, extracurriculars, recommendation style, interview polish, employment history, writing conventions, and even the confidence signaled by an applicant's manner are all shaped by prior distributions of support, training, and opportunity. Barocas and Selbst put the point in broader terms: data mining can discover "preexisting patterns of exclusion and inequality" and then treat them as useful regularities for decision-making (2016, p. 688). So,

when a model predicts future success from such features, it is not reading merit off a neutral surface but often formalising the afterlife of earlier selections.

O'Neil helps explain why this is not only a problem of biased historical data, but also of the form of modelling itself. Because every model simplifies, it necessarily omits context that may be crucial to the meaning of the variables it uses. In high-stakes screening, that omission allows a system to treat school pedigree, prior institutional affiliation, recommendation style, or résumé polish as tractable signals while bracketing the social conditions that produced them. In that way it can then present a stratified history as if it were an objective ranking, thereby laundering earlier inequality into present selection (O'Neil, 2016, pp. 6, 9–10).

That is one reason hiring and admissions are not interchangeable with more ordinary predictive contexts, because they operate as bottlenecks. In Fishkin's formulation bottlenecks are "the narrow places through which people must pass if they hope to reach a wide range of opportunities that fan out on the other side" (2014, pp. 1–2). This fits both university admissions and many labour-market screens. A university decision is not only about who enters one institution this year. It structures later access to networks, credentials, professional training, and labour-market pathways. Likewise, an early employment screen does not only sort applicants for one position. It often determines who gets the experience that later screens will treat as evidence of promise. In such domains, prediction is world-shaping: it does not simply track a future that will happen anyway but helps determine which futures become available to whom.

Once the bottleneck structure is visible, the engineer's defence looks even thinner, because 'accuracy-first' can present itself as modest only by ignoring the temporal structure of the decision. The claim might appear humble if all that is happening is that a model estimates who will succeed at a task, but that is not all the case. In a bottlenecked domain, the model participates in the creation of the very future it predicts. The selected candidates receive further advantages;

the rejected candidates do not. Prediction in such domains therefore does not merely describe opportunity structures but helps reproduce them.

Huq's analysis of algorithmic criminal justice makes the same temporal point in a more general way. He argues that the relevant question is not simply whether a tool satisfies some abstract fairness metric at a moment in time, but what its long-run dynamic effects are on racial stratification (2019, pp. 1045–1047, 1128–1129). The same lesson applies here, in screening domains, the relevant question is what kinds of social positions a model helps reproduce over time, not just whether it predicts well at a single moment.

Section 6: What Counts as a Solution?

Once the engineer's defence is rejected, the word solution becomes ambiguous: it can name a technical solution, in which the model is redesigned so that it no longer relies on protected status or its proxies; a governance solution, in which the system remains in use but is embedded in institutions that make its objectives, trade-offs, and grounds of decision contestable; or a structural solution, in which the role of the algorithm is sharply limited or eliminated altogether. These are not interchangeable, because a solution to proxy discrimination is not simply whatever produces a cleaner metric, but whatever prevents the system from using impermissible grounds to allocate opportunity.

The strongest ideal is technical elimination; if a system could isolate only the predictive contribution of non-suspect variables, without reconstructing protected status through less visible stand-ins, that would remove the sharpest form of the objection. Prince and Schwarcz's own menu of responses includes strategies of exactly this sort: restricting the use of non-approved factors, expanding the data available to separate suspect from non-suspect predictive power, and requiring 'ethical algorithms' that explicitly control for proxy discrimination (2020, p. 1258). In that sense, technical correction is the cleanest answer because it targets the mechanism directly,

but it is also often infeasible in practice. Proxy discrimination is difficult precisely because once directly predictive but legally suspect traits are excluded, AI systems may locate less intuitive stand-ins instead (Prince & Schwarcz, 2020, pp. 1263–1264). As Green argues, a system can satisfy a fairness metric while worsening injustice in practice, so technical elimination, though the strongest ideal, cannot be the normal or sufficient answer (2022, p. 18).

Governance is therefore the ordinary practical baseline. In high-stakes settings, the relevant issue is not only what prediction an algorithm generates, but what objective it has been built to optimise, what outcomes it is trained to predict, and what trade-offs are embedded in its design. The institutional response is to preserve the system's inputs, training data, objective, and decision rules in a form that makes them inspectable and contestable (Kleinberg et al., 2019, pp. 114–115). As Kroll et al. argue, the aim is accountability rather than naïve faith in source-code transparency alone (2017, pp. 657–660, 672–673).

Even governance is not always enough, because in bottlenecked domains, where decisions shape later life chances and inputs are already deeply structured by prior inequality, the right conclusion may be structural restriction or non-deployment. Green's positive proposal shifts the question from whether a decision rule satisfies some formal fairness criterion to whether and how algorithms can promote justice in practice (2022, pp. 24–25). Once the frame is widened in that way, some uses of prediction will no longer look reformable by better metrics or audit trails alone. If the domain is one in which the system's very function is to convert socially tainted signals into authoritative allocations, and if governance cannot secure permissible treatment, then limiting or refusing deployment is not a failure to solve the problem but the solution. This is especially plausible in hiring and admissions when an algorithm's predictive success depends on features whose meaning is inseparable from stratified opportunity.

Martin adds one final point: governance is not abstract institutional magic, because algorithms are value-laden and firms remain responsible for the role structure they design into

the decision process, including how much responsibility is left to human actors and how much is shifted into the system itself (2019, pp. 835–837). Non-deployment is not only a public-policy conclusion, but also sometimes the right conclusion for designers and deployers deciding whether a system should exist in its present form.

The hierarchy is straightforward: technical elimination is the strongest ideal, governance the ordinary practical baseline, and structural restriction or non-deployment appropriate where governance cannot secure permissible treatment.

Section 7: Conclusion

The engineer's defence begins from two truths: historical data are shaped by prior inequality, and engineers do not control the institutions that produced it; but those truths do not yield the conclusion the defence wants, and proxy discrimination shows why. Once protected status can be reconstructed through facially neutral features, the claim that a system did not use race, sex, disability, or other suspect traits no longer settles very much, because the issue is not formal exclusion alone but whether predictive success is being achieved through the informational value of protected status or its social residue (Johnson, 2021, pp. 9941–9943; Prince & Schwarcz, 2020, pp. 1260–1264).

That in turn exposes the defence's hidden premise that predictive validity is enough to make a variable permissible in high-stakes selection, and it is not. As Hellman shows, prediction bears first on belief, while fairness concerns what may be done to people; as Fishkin shows, merit cannot be defined by whatever predictors happen to sort well; and as Mayson shows, prediction in a stratified world imports the past into the future (Fishkin, 2014, p. 57; Hellman, 2020, p. 812; Mayson, 2019, p. 2218). In hiring and admissions, that problem is intensified because these are bottlenecked domains in which the inputs already reflect prior allocation and the output structures later opportunity (Fishkin, 2014, pp. 1–2).

The practical consequence is not that every algorithm must be abandoned, but that ‘solution’ must be disambiguated. Technical elimination is the strongest ideal, governance the ordinary practical baseline, and where governance cannot secure permissible treatment in bottlenecked domains, structural restriction or non-deployment is justified (Green, 2022, pp. 24–25). In high-stakes screening, an institution cannot justify algorithmic use merely by showing that a model predicts well; it must also show that the grounds on which it predicts are grounds it may permissibly act on.

Section 8: Bibliography

Ajunwa, I. (2020). The Paradox of Automation as Anti-Bias Intervention. *Cardozo Law Review*, *41*(5), 1671–1742.

Anderson, E. S. (1999). What Is the Point of Equality? *Ethics*, *109*(2), 287–337. https://doi.org/10.1086/233897

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671–732.

Fishkin, J. (2014). *Bottlenecks: A new theory of equal opportunity*. Oxford University Press.

Green, B. (2022). Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology*, *35*(4), 90-. https://doi.org/10.1007/s13347-022-00584-6

Hellman, D. (2020). Measuring Algorithmic Fairness. *Virginia Law Review*, *106*(4), 811–866.

Huq, A. (2019). Racial Equity in Algorithmic Criminal Justice. *Duke Law Journal*, *68*(6), 1043–1134.

Johnson, G. M. (2021). Algorithmic bias: On the implicit biases of social technology. *Synthese*, *198*(10), 9941–9961.

Kelly-Lyth, A. (2023). Algorithmic discrimination at work. *European Labour Law Journal*, *14*(2), 152–171. https://doi.org/10.1177/20319525231167300

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination In The Age Of Algorithms. *NBER Working Paper Series*, *w25548*.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, *165*(3), 633–705.

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, *160*(4), 835–850. https://doi.org/10.1007/s10551-018-3921-3

Mayson, S. G. (2019). Bias In, Bias Out. *The Yale Law Journal*, *128*(8), 2218–2300.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. The Crown Publishing Group. http://ebookcentral.proquest.com/lib/waterloo/detail.action?docID=6108230

Prince, A. E. R., & Schwarcz, D. (2020). Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa Law Review*, *105*(3), 1257–1319.

Tschantz, M. C. (2022). *What is Proxy Discrimination?* (arXiv:2205.05265). arXiv. https://doi.org/10.48550/arXiv.2205.05265