

Same System, Opposite Verdicts: Metric Discretion in AI Ethics Audits and the Limits of Disclosure

Shay Tsaban

shay.tsaban@gmail.com

Peres Academic Center

Research Article

Keywords: AI ethics audit, algorithmic accountability, alternative ethics measures, multiverse analysis, disclosure governance, alternative performance measures

Posted Date: August 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10576395/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Regulators increasingly require organisations to report a number for the fairness of an artificial intelligence system, and no standard fixes how that number is computed. This paper measures what that gap is worth. It introduces Alternative Ethics Measures, a construct modelled on the alternative performance measures that financial reporting has governed for three decades. Study 1 audits four decision systems, one deployed instrument and three canonical benchmarks, across the full range of defensible specifications, 91,572 in all. Within that range the verdict flips in all 7 pairings of a system with a protected attribute, the group identified as disadvantaged reverses, and the reported figure moves 6 to 30 times as far as sampling error. A validation exercise reproduces the error rates published in the COMPAS dispute and traces them to a reference population the original analysis never stated. Study 2 reads 6,797 published disclosure documents from two regimes, one mandatory and one voluntary, and finds only 38 reporting a quantified fairness measure. None justifies the selection, names the alternative not chosen, or reports a prior-period figure. Study 3 estimates what disclosure would recover. Reconciliation to a stated reference computation would close 78 per cent of the discretion window, adding consistency disclosure would bring the total to 82 per cent, and a measured residual would survive both. Governing the discretion therefore does not require settling which fairness metric is correct. Those choices can be made legible now.

JEL classification:M42, M48, K23, O33, M14

1. Introduction

In May 2016 ProPublica published an investigation of COMPAS, a proprietary instrument then scoring pretrial defendants in Broward County, Florida. Its reporters obtained risk scores for thousands of defendants, followed each case for two years, and computed the instrument's error rates by race. Among defendants who were not rearrested, Black defendants had been labelled high risk at far higher rates than white defendants (Angwin et al. 2016). That finding circulated as evidence that a machine used in criminal courts was biased.

Northpointe, the vendor, replied that the same scores satisfied predictive parity, since a given score carried the same probability of rearrest for Black and for white defendants (Dieterich et al. 2016). On that measure the instrument treated the groups alike. Two analyses of one system had examined the same defendants and produced opposite verdicts, and neither showed the other to have miscomputed a number.

The dispute was never about arithmetic. It turned on which measurement, computed over which population, should count as the finding. Both parties had answered with a defensible definition of fairness, and the definitions pointed in opposite directions. Neither publication stated the full set of choices behind its figures, and no standard existed that could settle the disagreement. That stalemate has never been broken.

Regulation has since made that choice unavoidable. The European Union's AI Act subjects high-risk systems to conformity assessment and post-market monitoring (European Union 2024). New York City requires employers using automated hiring tools to commission an annual bias audit and to publish the impact ratios it produces (Groves et al. 2024). Canada obliges federal institutions to complete and publish an algorithmic impact assessment before deploying an automated decision system (Treasury Board of Canada Secretariat 2023). Each regime demands a quantified finding, because a narrative judgement cannot be benchmarked against a threshold or compared across systems.

None of these instruments states which quantity the finding should be. An auditor working under any of them must select a metric family, fix a decision threshold, choose a reference population and decide which groups to compare. Published guidance specifies how systems are documented and accessed, and it is silent on each of these choices. The Broward County question, which computation counts, therefore travels with every high-risk deployment.

That impasse was not an accident of one case. Fairness admits many defensible operationalisations, and the central candidates cannot hold at once whenever base rates differ across groups (Kleinberg et al. 2017; Chouldechova 2017; Barocas et al. 2023). Which candidate ought to prevail is contested normative territory, and a decade of debate has not resolved it. What the debate has not produced is a magnitude, an estimate of how far the discretion moves the figure an audit reports.

Existing pathology constructs do not reach this case. Fairwashing describes a surrogate explanation engineered to conceal a discriminatory model from its auditor (Aïvodji et al. 2019, 2021). Audit-washing describes an audit process staged so that a self-assessment or a captured reviewer passes for an independent one (Goodman and Trehu 2022). Both locate the failure in the integrity of the audit or its process. An honest audit whose finding could have been materially different under another defensible specification sits outside both constructs, and no study has measured that possibility on a deployed system.

This paper names that residual discretion Alternative Ethics Measures and then measures it. The name follows the alternative performance measures of financial reporting, a domain that has governed a structurally similar discretion for three decades without settling which measure is correct (European Securities and Markets Authority 2015; Black et al. 2018). Readers of financial statements learned to treat an adjusted earnings figure as one defensible computation among several, because disclosure rules made the alternatives visible. Readers of quantified ethics claims have no comparable footing.

Three linked studies supply the measurement, and Fig. 1 summarises their design. Study 1 treats an ethics audit as a multiverse and enumerates every defensible specification of four decision systems, one deployed instrument and three canonical benchmarks, 91,572 specifications in all, spanning criminal justice, income classification, consumer credit and employment. Study 2 reads every quantified fairness claim published under two disclosure regimes in current operation and asks whether a reader could reproduce the figure disclosed. Study 3 conditions the Study 1 space on progressively fuller disclosure and quantifies how much of the range each disclosed element would recover.

The central finding can be stated at once. In all 7 system and attribute pairs examined, the verdict under the regulatory four-fifths rule flips within the space of defensible specifications. Even the group identified as disadvantaged reverses. None of these specifications misstates an input or departs from published guidance. Nor does any of them require bad faith. The movement also dwarfs the one uncertainty audits already report, running 6 to 30 times the size of sampling error. A published fairness figure on these systems is therefore a selection from a wide range, and nothing in current practice obliges the publisher to describe the range.

The other two studies locate the problem in practice and measure the remedy. Published disclosures, where they quantify fairness at all, document the computation and never the choice, so a reader can reconstruct what was measured and cannot tell what else could have been. Requiring auditors to state the choices they have already made would close most of the range at almost no cost. A measured residual would survive, and it marks what transparency alone cannot repair.

Four contributions result. The paper supplies what is, to the author's knowledge, the first measurement of audit-level metric discretion, taken on a deployed instrument and three canonical benchmark tasks. A decomposition then shows, with empirical weights, which of an auditor's choices moves the reported figure most. A reading of published disclosures establishes what current practice states and what it uniformly omits. Finally, a quantified bound is placed on how much of the discretion mandatory disclosure would remove and how much would remain.

Section 2 places the construct within the AI auditing, fairness measurement and financial reporting literatures and closes with four research questions. Section 3 outlines the design of the three studies, which Sections 4 to 6 then report in turn. Section 7 summarises the findings, develops their implications for theory, practice and policy, and weighs the limitations, and Section 8 concludes.

2 Background

2.1 From principles to quantified audits

The audit turn in AI ethics has a traceable origin. Ethics guidelines multiplied through the late 2010s and converged on the same abstract commitments while diverging almost completely on implementation (Jobin et al. 2019; Hagendorff 2020). Convergence at that level of abstraction left every operational question open. Principles alone cannot discipline conduct, because they lack the licensing and liability structures that established professions rely on (Mittelstadt 2019). Auditing emerged as the corrective, converting commitments into procedures that can be tested, repeated and compared (Mökander and Floridi 2021; Raji et al. 2020).

That turn now carries legal force. European law sets conformity assessment duties for high-risk systems through the AI Act (European Union 2024). New York City's Local Law 144 requires annual bias audits of automated employment decision tools and publication of the impact ratios they produce (Groves et al. 2024; Wright et al. 2024). Canada requires a published Algorithmic Impact Assessment for every federal

automated decision system (Treasury Board of Canada Secretariat 2023). Each regime demands a quantified finding, since a narrative judgement can be neither benchmarked against a threshold nor compared across systems.

An apparatus grew up to serve these obligations, and its lineage explains what it omits. Documentation standards were built to make provenance and subgroup performance legible to a reviewer (Gebru et al. 2021; Mitchell et al. 2019). Internal audit frameworks structure review before release, and external audit designs reach systems from the outside (Raji et al. 2020; Metaxa et al. 2021). Later work in the same line mapped the access auditors are granted and the thinness of the supporting ecosystem (Casper et al. 2024; Birhane et al. 2024; Raji et al. 2022; Ojewale et al. 2025).

Every instrument in that lineage specifies access and documentation. None specifies which quantified test becomes the reported finding. Published audits show the consequence, since most report a single quantity without stating why that quantity was chosen (Bandy 2021). The choice carries weight beyond the report, because publishing a figure changes vendor behaviour (Raji and Buolamwini 2019). Harmonised standards, through which the EU Act's conformity route runs, do not yet fix this measurement layer (Laux et al. 2024). Whether the audit turn changes conduct at all remains contested (Munn 2022; Rességuier and Rodrigues 2020), and the question pursued here survives either answer because it concerns what a quantified finding means once produced.

Practitioner studies record the same gap from the inside. Industry teams report that no guidance tells them which fairness criterion applies to their setting, and toolkit users often cannot justify the default their software selected (Holstein et al. 2019; Madaio et al. 2022; Deng et al. 2022). Organisational studies show these choices being made under product pressure and without clear ownership (Rakova et al. 2021). What the reported difficulty lacks is a magnitude, which is what this paper supplies. A magnitude also determines which of the discretion points defined below deserves governance attention first.

2.2 One question, many defensible measures

Metric plurality has its own lineage, and it begins in law rather than in computing. Translating a legal standard such as disparate impact into a statistical test requires choices about population and outcome that the standard does not dictate (Barocas and Selbst 2016). European non-discrimination law leaves the same space open, since it names no statistical criterion an automated decision must satisfy (Wachter et al. 2021). Neither body of law resolves the choice. A fairness question therefore reaches the auditor already unresolved.

Computing filled that space with many candidates and no canon (Dwork et al. 2012; Hardt et al. 2016). More than twenty definitions now circulate, each internally coherent and most computable from the same underlying data (Verma and Rubin 2018; Narayanan 2018; Castelnovo et al. 2022). Most descend from a psychometrics debate that ran for fifty years without resolution (Hutchinson and Mitchell 2019). Subsequent work has added definitions and retired none. Impossibility results then removed any

prospect of convergence, since calibration and error-rate balance cannot both hold when base rates differ across groups (Kleinberg et al. 2017; Chouldechova 2017).

What follows from the plurality is contested normative ground rather than a technical preference. Selecting a criterion encodes rival views of what makes a disparity wrongful, views traced to political philosophy and to the measurement step that turns a commitment into anything computable (Mitchell et al. 2021; Binns 2018; Hellman 2020; Jacobs and Wallach 2021). Formalising the question in any single test also discards contextual information the choice depends on (Selbst et al. 2019). Some argue that formal criteria cannot settle the question at all (Green 2022). An auditor inherits this disagreement intact, together with the obligation to report one number anyway.

This choice also bites empirically. Rankings of fairness-enhancing interventions depend on the criterion used to score them (Friedler et al. 2019). Clinical and public-sector models routinely satisfy one criterion while violating another (Pfohl et al. 2021; Rodolfa et al. 2020), and an audit surfaces only the disparities its chosen metric can register (Buolamwini and Gebru 2018; Obermeyer et al. 2019). Toolkits compute a menu of criteria and leave selection to the user, so a library default becomes a selection decision made at scale (Bellamy et al. 2019; Bird et al. 2020; Saleiro et al. 2018; Deng et al. 2022).

Each branch of this lineage ends in an argument about which criterion is right (Corbett-Davies et al. 2023). This paper asks a prior question and holds every candidate internally valid. What it measures is how far the reported number moves across the candidates, and whether a reader of a published figure can tell which one produced it. Neither quantity has, so far as the literature reviewed here shows, been measured on the audit of a deployed system. One precondition bounds the exercise, since every criterion examined requires a recorded protected attribute, and such attributes are often unavailable or inferred (Veale and Binns 2017; Andrus et al. 2021). Section 7.4 returns to that boundary.

2.3 Adjacent constructs: deception, multiplicity and the residual case

Three adjacent literatures approach the same terrain, and the space between them locates the present construct. One grew from deception. Fairwashing describes a surrogate explanation engineered to make an unfair model appear fair to its auditor (Aïvodji et al. 2019, 2021). Audit-washing extends the concern from the artefact to the process, where a captured or cosmetic audit lends legitimacy without scrutiny (Goodman and Trehu 2022; Costanza-Chock et al. 2022). Fairwashing presupposes a party who knows the true state and misrepresents it, and audit-washing reaches processes that are captured or merely cosmetic. In either form the research problem is detection rather than measurement.

A second literature reached the terrain from statistics. Predictive multiplicity showed that many near-optimal models assign conflicting predictions to the same cases (Marx et al. 2020; Watson-Daniels et al. 2023), and model multiplicity systematised the consequences for accountability (Black et al. 2022). Training randomness alone moves group fairness measures between runs of one pipeline (Ganesh et al. 2023; Cooper et al. 2024), and the construction of the measurement itself moves the measured disparity

(Lum et al. 2022). Fairness hacking then named the deliberate exploitation of metric plurality (Meding and Hagendorff 2024). Simson et al. (2024) brought multiverse analysis to the design, implementation and evaluation decisions of a single prediction pipeline, in order to make such exploitation visible.

The nearest of these designs share individual axes with the present one. Simson et al. (2024) vary the classification cutoff and the grouping used in evaluation within one modelling pipeline, and read the result on a single fairness measure for a benchmark prediction task. Meding and Hagendorff (2024) show that metric choice alone can reverse which group appears disadvantaged. What this paper adds is the audit as the unit of analysis. The system is held fixed and includes a deployed instrument, and the outcome is the regulatory verdict under a real threshold rather than the value of a metric. The axes an assessor controls, led by the choice of metric family and the reference population, are enumerated jointly, so their contributions can be weighed against each other and against sampling error. The enumeration is then joined to a corpus of published disclosures in Study 2 and to a quantified bound on the value of disclosure in Study 3, which no multiplicity or fairness-hacking study provides.

A third strand is the general critique of metric fixation, in which reliance on any measure invites optimisation pressure against the goal the measure stands for (Thomas and Uminsky 2022). That critique is broader than the present construct, which concerns the disclosure of one specific choice. The construct is broader than the deception constructs in turn, because it requires bad faith from nobody. An honest audit with full access can report a finding that another defensible specification would have reversed, and no existing construct measures how often that occurs.

2.4 The financial reporting precedent

Financial reporting met a structurally identical problem and left a documented trail. Generally Accepted Accounting Principles supply a canonical earnings baseline, yet public companies routinely disclose self-defined alternatives beside it. An alternative figure can carry information the standardised one does not (Bhattacharya et al. 2003; Lougee and Marquardt 2004). It is also managed opportunistically, with exclusions associated with meeting analyst forecasts and a widening divergence between street and GAAP earnings (Doyle et al. 2003, 2013; Bradshaw and Sloan 2002).

How this evidence accumulated matters for the analogy. Evidence of informativeness came early, so an outright ban would have destroyed something of value. Scrutiny then changed which items firms excluded rather than whether they excluded, and the alternative proved most informative exactly where the standardised figure was least so (Kolev et al. 2008; Leung and Veenman 2018). Regulation therefore arrived knowing that the discretion was both useful and manipulable, which is the position AI ethics measurement occupies now.

Regulators in this domain governed the discretion rather than banning it. European guidance requires issuers to define each alternative measure, reconcile it to the most comparable standardised figure, apply definitions consistently across periods and give the alternative no undue prominence (European Securities and Markets Authority 2015). Those four requirements operate on disclosure alone, and they

leave the substantive choice of measure with the firm. Their object is comparability between disclosures rather than correctness of any one measure.

ESG ratings replay the same anatomy for a business readership. Ratings of one firm diverge across raters, and the divergence decomposes into scope, measurement and weight choices, the counterparts of selection and definition here (Berg et al. 2022; Chatterji et al. 2016). Fuller disclosure does not mechanically produce convergence among raters (Christensen et al. 2022), which anticipates the ceiling Study 3 measures. Sustainability assurance likewise grew around quantities no authority had defined, and mandating such reporting achieves some of its aims and not others (Simnett et al. 2009; Christensen et al. 2021).

A decade of the alternative-measures regime also documents its limits. Comparability of non-GAAP earnings remains incomplete despite mature guidance, and informed opinion holds that the measures can inform or mislead depending on how they are constructed and disclosed (Black et al. 2018, 2021). Auditor involvement adds credibility that disclosure alone does not supply (Lamoreaux et al. 2026). The precedent therefore exhibits a working mechanism together with its ceiling.

Three elements of that mechanism transfer to AI ethics measurement: a definition requirement, a reconciliation requirement and a consistency requirement. One element does not transfer, because the European regime reconciles to a baseline that is itself canonical and legally mandated, and AI ethics measurement has no such baseline. Reconciliation here must instead mean reconciliation to a stated reference computation that the organisation selects and justifies. Study 3 measures what that substitution costs.

2.5 Alternative Ethics Measures and the research questions

The strands above converge on a construct that none of them names. Alternative Ethics Measures are quantitative indicators of an AI system's ethical performance that an organisation selects, operationally defines, computes and discloses. The definition applies only where no legally mandated or professionally canonical measurement standard governs the dimension concerned. Fairness supplies the test case throughout, but the definition covers any ethical dimension that meets its conditions. Its name marks the parallel with alternative performance measures deliberately, since Section 2.4 shows that the governance vocabulary transfers.

Three features do the definitional work. An indicator qualifies when it is quantitative rather than narrative, so that it can be benchmarked and compared. The organisation must retain discretion over selection, definition and disclosure. And the discretion must persist because no canonical baseline has been mandated. That residual quality confines the construct to dimensions a standard does not yet govern, and it dissolves wherever one is adopted.

Discretion enters the reported figure at three identifiable points in the audit process, each observable without access to intent:

1. Selection, the choice of which legitimate metric family becomes the reported finding.
2. Definition, the choice of threshold, reference population and comparison rule within that family.
3. Consistency, whether these choices change across reporting periods and whether any change is disclosed.

These are points in a process rather than types of motive. The same selection can be made in good faith or opportunistically, and governing the point does not require inferring which, since a documented process can be inspected where a motive cannot. Each point is observable in a published report, or observable by its absence, which is what Study 2 exploits. Each can also be weighted empirically by how much of the reported figure it moves, which is what Study 1 does. Four questions organise the measurement, and the three studies answer them in order.

How wide is the range of fairness figures a competent auditor could defensibly report for the same system, and does the range cross the regulatory threshold? Which discretion point moves the figure most, and how does the movement compare with sampling error? Do published quantified fairness disclosures state the choices a reader would need to reproduce them? And how much of the range would mandatory disclosure remove, and how much would remain?

3 Overview of the studies

The four questions closing Section 2 are addressed in three linked studies, and Fig. 1 shows the design. Study 1 treats the audit of 4 decision systems as a multiverse and enumerates every defensible specification, which answers the first two questions, on the width of the range and on where the movement lies. Study 2 answers the third by reading every quantified fairness claim published under two disclosure regimes in current operation. Study 3 answers the fourth by returning to the enumerated space and conditioning it on what an auditor has disclosed.

All three studies share one measurement scale, so the remedy tested in Study 3 is evaluated on the same space in which Study 1 measures the problem. All data are public. One analysis pipeline downloads each source, enumerates the specification space, codes the disclosure corpus and reproduces every table and figure from it. No result depends on access to any proprietary record.

4 Study 1: The discretion window in four audited systems

4.1 Design

An AI ethics audit is treated here as a multiverse. Each combination of defensible choices is one specification, and each specification yields one reported figure (Steege et al. 2016; Simonsohn et al. 2020). The design commitment is to enumerate that space in full rather than to defend any single

specification. Enumeration records every finding the system could honestly have produced, which is the width the first question asks about.

The unit of analysis is therefore a specification, not a system. Each system, its data and its outcomes are held fixed, and only the choices an assessor makes after the system exists are varied. Holding the system fixed isolates the movement in the reported figure that is attributable to the audit rather than to the model or the sample. Section 4.3 defines the axes along which the choices vary.

Evidence from other fields shows the concern is not hypothetical. Independent teams analysing identical data reach materially different conclusions across psychology, neuroimaging and sociology (Silberzahn et al. 2018; Botvinik-Nezer et al. 2020; Breznau et al. 2022). That evidence grew from an older worry about undisclosed analytic flexibility within single studies (Simmons et al. 2011; Gelman and Loken 2013). Machine learning has a parallel record, with much published work documented too thinly to rerun (Gundersen and Kjensmo 2018; Pineau et al. 2021).

A research finding is offered to a discipline that can reanalyse it. An audit reports one number to a regulator against a threshold, so undisclosed flexibility carries a higher price there. Clinical research involving artificial intelligence now has reporting standards that specify what a published study must state (Collins et al. 2024; Liu et al. 2020). No comparable standard governs what an AI ethics audit must state. Study 1 measures what that absence leaves open.

4.2 Systems and data

Study 1 audits 4 decision systems, and Table 1 lists each with its sample and protected attributes. Two grounds fixed the selection. Each system sits in a sector the auditing literature treats as high stakes, spanning criminal justice, income classification, consumer credit and employment. Each also rests on public data, so every figure reported below can be recomputed from source. The first system is COMPAS, the pretrial risk instrument used in Broward County, Florida, on records released by ProPublica (Angwin et al. 2016).

Table 1
The four decision systems audited

System	Sector	Outcome audited	Protected attributes	Specifications	Source
COMPAS pretrial risk (Broward County)	Criminal justice	Two-year recidivism	race, sex	6,117	ProPublica
Census income classification (Adult)	Income	Annual income above \$50,000	race, sex	4,996	UCI Machine Learning Repository
German consumer credit (Statlog)	Consumer credit	Good credit risk	sex	556	UCI Machine Learning Repository
ACS employment prediction (3 states, 2015–2019)	Employment	Employment status	race, sex	79,903	US Census Bureau
Note. Every source is public. Specification counts are for the defensible core defined in Section 4.4.					

The deployed proprietary decile score is audited directly, alongside models fitted on the same features, so at least one audited system is a real instrument that was in real use. The remaining three are established benchmarks. Adult is the census income extract long used to study fair classification, obtained from the UCI repository. German Credit comprises one thousand consumer credit applications from the Statlog collection, classified as good or bad risk. Benchmarks are retained because the fairness literature and its toolkits evaluate on them, so their defensible choices are documented rather than invented here. Treating these datasets as grounds for substantive fairness conclusions has been criticised (Bao et al. 2021). They serve in this study to measure audit-level measurement discretion, not to support claims about the underlying populations.

ACS Employment predicts employment status from American Community Survey microdata, following the task definition of Ding et al. (2021), for three states across five survey years. The panel supplies the temporal variation the consistency axis requires and the study’s largest sample. Protected attributes are used as the source records state them, which yields 7 system and attribute pairs. Pairs rather than systems are the unit at which results are reported, since a system with two recorded attributes supports two distinct audits.

4.3 The specification space

Table 2 lists the seven axes of the specification space, the discretion point each operationalises and the levels each takes. A specification fixes one level on every axis, and the axes map onto the discretion points defined in Section 2.5. Selection is represented by the metric family axis. Eight families are computed, spanning comparisons of selection rates, error rates, predictive values and accuracy across

groups. Each family is individually coherent and appears in deployed audit practice or in the regulatory record, so none can be excluded as indefensible.

Table 2
The specification space: axes, discretion points and levels

Axis	Discretion point	Levels
Metric family	Selection	8 families, each comparing a group-level quantity across groups: statistical parity (selection rate), equal opportunity (true positive rate), false positive rate parity, false negative rate parity, predictive parity (positive predictive value), negative predictive value parity, accuracy parity, true negative rate parity
Decision threshold	Definition	5 rules: fixed cut at 0.5, cut matching the base rate, cut maximising Youden's J, top decile, top quartile
Reference population	Definition	3 defensible population filters per system
Group comparison rule	Definition	4 conventions: canonical named pair, aggregated minority against the reference category, two largest groups, widest disparity
Reporting period	Consistency	5 survey years by 3 states (census employment panel)
Model class	Benchmark	logistic regression, gradient boosting, random forest
Resampling seed	Benchmark	3 independent draws
<i>Note.</i> The two benchmark axes measure variation in the system and in the sample rather than in the audit, and give the discretion axes a scale to be judged against. For COMPAS the deployed decile score, rescaled to the unit interval, enters the model-class axis as a fourth level.		

Definition discretion enters through three axes. Five threshold rules convert a score into a decision, running from a fixed cut at 0.5 to top-share rules. Three reference populations per system each correspond to a filter an auditor could defend, such as the screening filter described in ProPublica's methodology or a felony-only restriction for COMPAS. Four comparison rules fix which groups the ratio runs between where an attribute has more than two levels, and only the named pair applies where it is binary.

Consistency discretion is carried by the reporting period axis, which crosses five survey years with three states on the census panel. Two further axes serve as benchmarks rather than as discretion. Model class varies across three learners and captures variation in the system rather than in the audit of it. Resampling seed varies across three draws and captures sampling error, giving the window a scale to be judged against. No auditor claims a figure is precise beyond its sampling error, so this benchmark anchors the comparison in Section 4.7.

Every reported figure is an impact ratio, the less favoured group's rate over the more favoured group's, which by construction never exceeds one. This is the form the four-fifths rule of the Uniform Guidelines on Employee Selection Procedures assumes (Equal Employment Opportunity Commission 1978), and the form in which the fairness literature operationalised disparate impact (Feldman et al. 2015). New York City's bias audit regime and hiring audit practice use the same form (Raghavan et al. 2020; Wilson et al. 2021). Applying the 0.80 threshold uniformly across metric families makes verdicts comparable across the space. The rule's legal home is narrower, and Section 4.9 restricts the analysis to that scope.

4.4 The defensible core and outcome measures

A specification enters the primary analysis only if each compared group holds at least 100 cases and the rate being compared rests on at least 30. Both constraints were fixed before any result was examined. Without them the space would admit arithmetically valid but professionally indefensible cells, such as a predictive value computed over a handful of selected cases, and the window would partly be an artefact of those cells. Section 4.9 reports how the headline figures move when the boundary is drawn differently. Four outcome measures then summarise each system and attribute pair:

1. The discretion window is the spread of the impact ratio across specifications, given as the full range, the interquartile range and the fifth to ninety-fifth percentile range, so no headline rests on a single extreme cell.
2. Verdict instability records whether both a pass and a fail against the four-fifths rule occur within the space.
3. Direction reversal records whether the group identified as disadvantaged changes across specifications.
4. Marginal variance shares record how much of the variance in the reported figure lies between the levels of each axis, and they are marginal because the design is crossed rather than orthogonal.

One framing point governs the reporting. The specifications enumerated here are not a sample from a population of specifications, since the space is defined in advance and traversed in full. Ranges are therefore reported without confidence intervals. Sampling uncertainty is real and is handled separately, by the seed axis, which measures how far a fixed specification moves when the data beneath it are redrawn. That axis carries the whole of the inferential load in Study 1.

4.5 The width of the window

The same system supports almost any reported figure. Enumerating the defensible core produced 91,572 specifications from 483 fitted model cells, spanning 4 systems and 7 system and attribute pairs. Table 3 reports the window for each pair, and Fig. 2 plots the full specification curve for four of them. Across pairs the reported impact ratio spans a median of 0.94 on a scale bounded at one. Even the narrowest window reaches 0.79. The widest covers the scale entirely at 1.00.

Table 3
The discretion window, by system and protected attribute

System	Attribute	Specs	Median	Range	IQR	P5-P95	% pass	Flips	Reverses
COMPAS	race	4,717	0.78	0.95	0.34	0.61	47.3	Yes	Yes
COMPAS	sex	1,400	0.86	0.94	0.26	0.66	62.6	Yes	Yes
Adult income	race	3,916	0.91	0.92	0.34	0.65	68.9	Yes	Yes
Adult income	sex	1,080	0.86	0.89	0.48	0.74	62.2	Yes	Yes
German credit	sex	556	0.87	0.79	0.17	0.40	70.1	Yes	Yes
ACS employment	race	63,807	0.90	0.99	0.21	0.58	69.3	Yes	Yes
ACS employment	sex	16,096	0.88	1.00	0.20	0.84	67.9	Yes	Yes

Note. The impact ratio is the less favoured group's rate as a fraction of the more favoured group's, so it cannot exceed one. Pass is against the four-fifths rule. Verdict flips records whether both a pass and a fail occur within the space; direction reverses records whether the group identified as disadvantaged changes.

Robust measures tell the same story. A range can always rest on one extreme cell, so two resistant measures accompany it. Median interquartile range across pairs is 0.26, and the median fifth to ninety-fifth percentile range is 0.65. Every spread measure is computed within a single system and attribute pair, so no pooling across systems inflates it. Even the middle half of a typical pair's space covers a substantial share of a scale whose whole length is one. Width of this order comes from the body of the distribution rather than from its tails.

Put in the terms an auditor would use, an organisation auditing five of these seven pairs could report an impact ratio near 0.00 or near 1.00, and even the narrowest window, German credit at 0.79, spans most of the scale. Either endpoint rests on a specification satisfying the inclusion rules fixed in advance, produced without misstating a single input and without departing from any published guidance. What the report says therefore depends more on the specification chosen than on anything the system did.

4.6 Verdict instability and direction reversal

Movement of this width crosses the line regulators apply in practice, and it does so in every pair examined. Both a pass and a fail against the four-fifths rule occur within the space in 7 of the 7 system and attribute pairs. Across pairs the share of specifications that pass runs from 47.3 to 70.1 per cent, with a median of 67.9 per cent. At the low end of that distribution the verdict approaches a coin toss, and

even at the median one specification in three fails. The choices that settle it never appear in the audit report.

A stronger result follows. Under different defensible specifications, the identity of the disadvantaged group reverses in 7 of 7 pairs. That metric choice can flip the direction of an apparent bias has been shown before on a single dataset (Meding and Hagendorff 2024). The enumeration shows the flip is systematic rather than exotic. Nor is it an artefact of near-parity cells. In all 7 of 7 pairs the space contains specifications placing each group below the 0.80 line, and the minimum reported ratio in each direction reaches 0.24 or lower. A reversal changes which group an audit says is being harmed, and it turns on choices that current disclosure practice leaves unstated.

Instability on this scale gives the window of Section 4.5 its full practical meaning. A wide range matters only where it crosses a threshold carrying institutional force, and here it crosses the operative threshold in every case examined. An organisation relying on any of these systems could obtain a passing audit or a failing one from the same records, with every party acting in good faith. Nothing in the published report format would distinguish the two outcomes.

4.7 Where the movement comes from

Selection dominates the movement. Table 4 and Fig. 3 decompose the variance in the reported impact ratio across the axes of the space. Metric family alone accounts for between 14 and 94 per cent of that variance, with a median across pairs of 45 per cent, the largest single share in every system examined. It is also the choice a published audit leaves least visible, since a report naming one family shows nothing of what the other families would have found.

Table 4
Share of variance in the reported impact ratio attributable to each axis

System	Attribute	Metric	Threshold	Population	Group rule	Period	Model	Seed
COMPAS	race	72.4	0.8	0.6	3.5		0.5	0.1
COMPAS	sex	44.7	1.8	0.2			13.7	0.5
Adult income	race	87.9	0.0	0.0	6.3		0.1	0.0
Adult income	sex	94.4	0.4	0.0			0.3	0.1
German credit	sex	14.1	5.8	1.9			0.3	1.7
ACS employment	race	22.7	6.5	1.4	15.4	2.4	0.2	0.0
ACS employment	sex	31.1	17.5	2.0		0.7	0.1	0.0

Note. Columns are the axes of the specification space: metric family (selection), threshold, reference population and group comparison rule (definition), reporting period (consistency), and model class and resampling seed (benchmarks rather than discretion). Marginal share of variance in the reported impact ratio lying between levels of each axis, in per cent. The design is crossed rather than orthogonal, so shares do not sum to 100. Blank cells indicate an axis with a single level for that pair. Each share is an eta squared, the variance of cell-count-weighted level means over the total variance, computed axis by axis. Main effects need not exhaust the variance, and the remainder lies in interactions and residual variation.

Definition choices come second and are not negligible, accounting for between 0 and 23 per cent in total. Within that group the decision threshold and the comparison rule usually carry more weight than the reference population, although Section 4.8 shows a reference population moving a headline figure on its own. Reporting period contributes at most 2.4 per cent. The resampling seed contributes at most 1.7 per cent. Because the design is crossed rather than orthogonal, the shares are marginal and do not form a partition summing to one.

These weights also settle a question the construct could not settle by stipulation. Selection, definition and consistency move the reported figure by materially different amounts, in a stable order across four unrelated systems, so the taxonomy of Section 2.5 earns its distinctions empirically. The benchmark axes stay small in six of the seven pairs. The exception is model class for COMPAS on sex, at 13.7 per cent, and the seed never exceeds 1.7 per cent anywhere. With that one exception, movement in the reported figure is attributable to how the system was audited rather than to how it was built or sampled.

Discretion also dwarfs the one uncertainty audits currently acknowledge, which is sampling error. Holding every specification choice fixed and redrawing the data beneath it moves the reported impact ratio by an average of 0.07. Letting the specification vary moves the figure by 0.93. Across pairs the ratio

between the two movements runs from 6 to 30 to one, with a median of 17, and Table 5 reports it for each pair. One asymmetry belongs in the reading. The specification range spans thousands of cells while the seed range rests on three redraws of each cell, so the multiple is an order-of-magnitude statement rather than a precise constant. The uncertainty that current practice quantifies and discloses is an order of magnitude smaller than the uncertainty it leaves undisclosed.

Table 5
Discretion measured against sampling error

System	Attribute	Full window	Sampling error only	Ratio
COMPAS	race	0.95	0.075	12.6
COMPAS	sex	0.94	0.113	8.3
Adult income	race	0.92	0.044	20.8
Adult income	sex	0.89	0.038	23.7
German credit	sex	0.79	0.134	5.9
ACS employment	race	0.99	0.059	17.0
ACS employment	sex	1.00	0.033	30.3
<i>Note.</i> Full window is the range of the reported impact ratio across all specifications. Sampling error only is the mean range across specifications that agree on every other axis and differ only in the resampling seed.				

4.8 Validation: reproducing the COMPAS dispute

Results so far rest on a space this study constructed, so an outside benchmark is needed. One case supplies it. The COMPAS controversy saw two parties publish contradictory quantified fairness findings about the same deployed instrument on the same defendants. ProPublica reported error rates that differed sharply by race (Angwin et al. 2016; Larson et al. 2016), and the vendor replied that the instrument satisfied predictive parity and accuracy equity (Dieterich et al. 2016). Both sets of figures sit in the public record. Two checks were run on the deployed decile scores rather than on any reconstruction. The decile score is rescaled to the unit interval before the threshold rules of Section 4.3 apply, and the deployed instrument enters the space as a fourth level of the model axis.

The first check recovers the published figures, and the recovery itself surfaces an undisclosed choice. Computed on the unfiltered file of 7,214 defendants, the recomputed false positive rates are 44.8 per cent for Black defendants and 23.5 per cent for white defendants. That reproduces all four published error rates to a total deviation of 0.01 percentage points. Computed on the 6,172 defendants remaining after the screening and charge filter the same methodology describes, the figures move by between 0.5 and 2.5 percentage points.

Nothing recovered here is an error. Both computations are correct arithmetic on the same instrument and the same outcome, and both populations are defensible readings of the same published methodology. The construction of the released file has been examined before (Barenstein 2019). The recovery adds a narrower point. The published table is consistent with two populations the same methodology describes, the article does not state which produced it, and a reader cannot tell from the disclosure alone. Verifying the figures required computing them twice, once per candidate population, and the second computation is what identified the choice.

The second check places both parties inside the enumerated space. Restricted to the 1,200 specifications computed on the deployed instrument, 2 per cent of false positive rate parity specifications pass the four-fifths rule, against 83 per cent of predictive parity specifications. The mean absolute within-decile recidivism gap between Black and white defendants is 4.0 percentage points. Predictive parity is the impact-ratio family closest to the vendor's defence, and this calibration-style gap is reported alongside the enumerated space rather than within it (Chouldechova 2017). The two published positions are two points in one specification space, and the dispute persisted because nothing in the record fixed which point either party had chosen.

The validation carries two conclusions for the framework. It reproduces independently published figures on a deployed instrument to within rounding, which is the accuracy a verification exercise demands. And it shows the field's most cited algorithmic audit instantiating the definition discretion this paper measures, in the direction the framework predicts. A dispute long treated as a contradiction to be adjudicated resolves into an unstated specification choice.

4.9 Sensitivity and scope

The window belongs to the space, not to its boundary. Table 6 recomputes the headline figures under four alternative definitions of the defensible core, and none overturns the conclusion that the window crosses the regulatory threshold. Removing the widest-disparity comparison rule, the most adversarial convention, narrows the median range from 0.94 to 0.89. Tightening the denominator requirement from 30 to 100 trims it to 0.92, and readmitting every arithmetically valid cell leaves it unchanged at the printed precision.

Table 6
Sensitivity of the headline figures to the definition of the specification space

Specification space	Specs	Median range	Median P5-P95	Median IQR	% pairs verdict unstable	% window closed by reconciliation
all arithmetically valid specifications	92,779	0.94	0.65	0.25	100	77
defensible core (denominator > = 30)	91,572	0.94	0.65	0.26	100	78
core, denominator > = 100	87,562	0.92	0.65	0.27	100	78
core, excluding the widest-disparity group rule	74,422	0.89	0.58	0.26	100	78
core, four most widely used metric families only	45,103	0.94	0.74	0.38	100	71
<i>Note.</i> Each row recomputes the headline quantities over a different definition of the specification space. The second row is the pre-specified primary analysis reported throughout.						

Restricting to the four most widely used metric families is the most conservative of the variants, and verdict instability survives it in every pair. Scope is nevertheless the sharpest objection the design invites, because in its legal home the four-fifths rule is a selection-rate test (Equal Employment Opportunity Commission 1978). Applying one threshold across eight families makes verdicts comparable, and it also reaches metrics the rule was never written for.

Table 7 therefore measures the scope objection directly, restricting the space to selection-rate metrics alone. That restriction removes the selection axis entirely, leaving the threshold, the reference population and the comparison rule as the only remaining audit choices, alongside the benchmark axes. Within the narrowed space the verdict still flips in four of the seven pairs, and the disadvantaged group still reverses in five. Median window falls from 0.94 to 0.72. Widening slightly to the three rate-based families in ordinary audit use restores instability in all seven pairs.

Table 7
Verdict instability when the space is narrowed to the four-fifths rule's own scope

Metric families in the space	Pairs	Median range	Median IQR	% pairs verdict unstable	% pairs direction reverses
all eight metric families	7	0.94	0.26	100	100
selection-rate metrics only (the rule's legal home)	7	0.72	0.20	57	71
rate-based metrics (selection, TPR, FPR)	7	0.92	0.38	100	100
<i>Note.</i> The four-fifths rule is a selection-rate test in its legal home. Each row narrows the specification space to the metric families named and recomputes the headline quantities over what remains.					

An honest reading carries a qualification and a remainder. Part of the instability reported in Section 4.6 does come from spanning metric families a single legal test was never meant to span. What remains still moves most of these systems across the regulatory line with the metric family already fixed. Definition choices alone, made after the metric is settled, are enough to change what an audit reports to a regulator.

5 Study 2: What published disclosures contain

5.1 Two disclosure regimes

Study 2 turns from what an audit could report to what published disclosures state in practice. It reads two disclosure regimes in full, chosen because both consist of real documents about systems in real use rather than research artefacts. Together they contribute 6,797 documents. Each document was examined for any quantified fairness claim and for whether a reader could reproduce the figure claimed.

The first regime is mandatory and small. Under Canada's Directive on Automated Decision-Making, a federal institution must complete an Algorithmic Impact Assessment before deploying an automated decision system and publish it on the open government portal (Treasury Board of Canada Secretariat 2023). Every published assessment was retrieved, giving 25 documents that cover live administrative systems in immigration, taxation, employment and benefits.

The second regime is voluntary and far larger. Model cards were proposed so that performance and bias could be reported per subgroup rather than in aggregate, which makes the genre's own standard the natural benchmark (Mitchell et al. 2019). Cards were retrieved for the most downloaded models on Hugging Face, giving 6,772 documents with substantive card text. Transparency scoring of the model ecosystem finds evaluation disclosure among its weakest dimensions (Bommasani et al. 2023), and a systematic reading of 32,111 cards finds the evaluation and limitations sections among the least

completed (Liang et al. 2024). Those studies count what a card contains. Study 2 asks whether a quantified fairness figure, where one exists, can be reproduced from what surrounds it.

New York City's bias audit regime would have been the most direct source, since it mandates publication of impact ratios by independent auditors. No central registry of those audits exists, and a systematic collection effort found that only a small minority of covered employers had posted one (Wright et al. 2024). Any corpus assembled from whatever can be found would be selected on willingness to publish, which is the variable at issue, so none was built.

5.2 Coding procedure

A corpus of 6,797 documents is too large to read and proved too heterogeneous to code automatically. Early automated coding erred in both directions across several iterations, so the final design gives each stage to the method that performs it dependably. An automated screen carries the volume and human reading carries the judgement. In consequence the screen fixes the denominator and the manual pass fixes the numerator.

Stage one is a deliberately generous screen tuned for recall rather than precision. It flags any document in which a fairness-measure term, a named bias benchmark, a subgroup performance comparison or a bias evaluation heading appears within 400 characters of a rate-like number. Over-collection is the intended behaviour, because a candidate wrongly flagged costs one reading while a disclosure wrongly missed is never seen at all. The screen's recall over unflagged documents is not directly estimated, a bound Section 7.4 records.

Stage two reads every flagged candidate against a single question, whether the document reports a named fairness or bias measure with a numeric value for the model or system concerned. Only hand-confirmed documents enter the analysis. Each judgement is recorded with its reason and distributed with the analysis code, so any decision can be inspected and contested. Borderline cases were resolved against the paper's own argument.

Confirmed disclosures were then coded on the specification elements Study 1 shows determine the reported figure. Three elements apply to every document: whether the measure is named, whether the evaluation population is stated, and whether the compared groups are identified. A decision threshold is well defined for a classification audit and often meaningless for a generative-model benchmark, which most of this corpus contains. Threshold disclosure is therefore reported separately, so the primary standard never penalises a document for omitting an inapplicable element. Both the screen and every judgement precede the results reported below.

5.3 What the documents disclose

Quantified fairness disclosure is rare almost to the point of absence. The screen flagged 76 candidates from the 6,797 documents, and manual reading confirmed 38 while rejecting 38. Confirmed disclosures

therefore amount to 0.56 per cent of the corpus, and the rejection share near one half is the expected signature of a screen built for recall. Table 8 reports the full breakdown.

Table 8
Disclosure of the specification behind confirmed quantified fairness measures

Specification element	Disclosures	Share (%)
Fairness measure named	38	100.0
Evaluation population stated	26	68.4
Compared groups identified	27	71.1
All three that always apply	26	68.4
Decision threshold stated	0	0.0
Alternative measure named or choice justified	0	0.0
Prior-period figure reported	0	0.0
Independent attestation claimed	0	0.0
<i>Note.</i> Base is the 38 documents confirmed by manual reading to report a named fairness measure with a value, out of 76 flagged by a generous automated screen over 6,797 documents. A decision threshold is not meaningful for every measure in this corpus, so the primary standard is the three elements that always apply.		

Even that count overstates the practice. The 38 confirmed documents resolve to 9 distinct underlying disclosures, the remainder being quantisations, re-uploads and forks that carry the original card text unchanged. In substance the corpus contains 9 organisations that disclose a quantified fairness figure. Both counts are informative, since the first describes the documentation a practitioner encounters and the second the disclosure practice behind it.

The mandated regime contributes nothing to the confirmed set. None of the 25 Algorithmic Impact Assessments reports a named fairness measure with a value, although every one covers a system in live administrative use. That absence is a property of the instrument rather than non-compliance, because the assessment collects structured responses about risk and mitigation without requiring a number on any ethical dimension. The discretion measured in Study 1 cannot arise in these documents because the measurement itself never arises.

Where a quantified claim does exist, the computation behind it is documented with some care. An evaluation population is stated in 26 of the confirmed disclosures (68.4 per cent) and the compared groups are identified in 27 (71.1 per cent). 26 disclosures (68.4 per cent) state all three always-applicable elements, a better record than the framework anticipated. The measure itself is named in every confirmed document, which follows from how confirmation was defined.

The disclosures satisfying all three concentrate in a small number of carefully written cards. One family reports word error rate by gender and by age band with the number of utterances in each group. Another reports classification accuracy across race and gender categories on a named evaluation dataset. A third gives a representation figure computed on a stated sample, so an organisation that quantifies at all tends to say what was measured, on what, and between which groups.

The choice behind the computation is documented nowhere. Of the confirmed disclosures, none states a decision threshold, reports a prior-period figure, claims independent attestation, or names the alternative measure not chosen. That last omission is the costliest, since metric family accounts for more of the variation in the reported figure than every other axis combined (Section 4.7). A reader of these documents can often reconstruct what was measured. No reader can establish what else could have been measured or what it would have shown.

6 Study 3: What disclosure would recover

6.1 Logic of the test

Study 3 tests the remedy the financial reporting precedent recommends. Disclosure does not remove discretion. It removes the reader's uncertainty about which choice an auditor made, and that reduction has a measurable value inside the enumerated space of Study 1. Measured this way, the question is how far the range of figures consistent with a published report narrows as elements are added to it.

The conditioning logic is simple. A reader told the metric family can discard every specification that uses another family. Adding the threshold, the reference population and the comparison rule discards more, so the set of figures consistent with the disclosure shrinks at each step. What remains after each step measures its worth, and the verdict ambiguity remaining measures whether the residual still matters for the regulatory outcome. Because the space is enumerated in full, each restriction is computed directly rather than assumed.

6.2 The return on each disclosure

Metric family is the single most valuable disclosure. With nothing disclosed, the reported figure could sit anywhere in a range averaging 0.93. Disclosing the family alone cuts the range to 0.52 and closes 44 per cent of the window, consistent with the variance decomposition in Section 4.7. Table 9 reports the progression and Fig. 4 traces it across the space, pooled over systems with each pair weighted equally.

Table 9
The discretion window under progressive disclosure

Disclosed by the auditor	Residual range of the reported figure	% of window closed	% of cells with an ambiguous verdict
nothing disclosed	0.93	0	100
+ metric family	0.52	44	80
+ decision threshold	0.32	65	55
+ reference population	0.27	72	44
+ group comparison rule (full reconciliation)	0.21	78	35
+ reporting period (consistency)	0.16	82	27
+ model class (system fixed; sampling error only)	0.07	92	12
<i>Note.</i> Pooled across systems, each weighted equally. Each row reports the range of reported figures still consistent with everything disclosed to that point. Elements are disclosed in the stated order, so each row's reduction is conditional on the rows above. The order follows the variance decomposition in Table 4.			

Full reconciliation extends the gain substantially. Adding the threshold, the reference population and the comparison rule, the elements that together constitute reconciliation to a stated reference computation, leaves a residual range of 0.21 and closes 78 per cent of the window. Consistency disclosure of the reporting period brings the total to 82 per cent. No document in Study 2 reports a prior-period figure, so the consistency mechanism is at present exercised nowhere.

Verdict ambiguity falls along the same path. With nothing disclosed, every conditioning cell contains both a pass and a fail against the four-fifths rule, so a disclosure regime inherits the full instability of Section 4.6. After full reconciliation 35 per cent of cells still do, after consistency disclosure 27 per cent, and once the model class is also fixed, so that only sampling error remains, 12 per cent.

The first conclusion is that reconciliation is cheap as well as effective. It demands no new measurement, no canonical standard and no ruling on which fairness metric is correct. An organisation need only record four choices it has already made. On the evidence here that record removes roughly three quarters of the range within which its reported figure could otherwise sit.

The second conclusion carries equal weight, because the progression also measures a ceiling. A residual range of 0.21 survives full reconciliation, and 0.16 survives consistency disclosure. What remains at that point is variation the named disclosures cannot reach. The reporting period exists as a choice only on the census panel, the model class is an audit choice only where the auditor fits the model rather than scores a deployed instrument, and beneath both sits sampling error, which no disclosure

removes. Even with the model class fixed, 12 per cent of cells still contain both verdicts. The residual is a result in its own right rather than a qualification of the remedy. It marks the boundary between what disclosure regulation can deliver, most of the window at the cost of recording four choices, and a remainder that only a substantive measurement standard or tighter estimation could narrow.

7 Discussion

7.1 Summary of the findings

The four questions of Section 2.5 are answered in turn. The first asked how wide the defensible range is. It spans most of the scale and crosses the regulatory line in every case examined. Across 91,572 specifications the reported impact ratio moves by a median of 0.94 on a scale bounded at one. A pass and a fail both occur in 7 of 7 system and attribute pairs, and the disadvantaged group reverses in 7. Restricted to selection-rate metrics, the four-fifths rule's legal home, the verdict still flips in four of seven pairs. A COMPAS check anchors this to the public record, reproducing ProPublica's error rates to 0.01 percentage points.

The second question asked where the movement comes from, and selection dominates it. Metric family alone accounts for 14 to 94 per cent of variance, with a median of 45 per cent. Definition axes together contribute 0 to 23 per cent, the period at most 2.4 and the seed at most 1.7. Varying the specification moves the Fig. 6 to 30 times as far as redrawing the data, with a median ratio of 17. Sampling error, the one uncertainty audits routinely disclose, is the smaller quantity in every pair.

The third question asked what published disclosures reveal. They describe the chosen computation and stay silent on the choice. Of 6,797 documents, 38 report a named fairness measure with a value, or 0.56 per cent of the corpus. These resolve to 9 distinct disclosures, all from the voluntary regime. None of the 25 mandated impact assessments contains one. Where quantification exists, 68.4 per cent state the measure, the evaluation population and the compared groups. None states a threshold, a prior-period figure, an attestation claim or the alternative measure not chosen.

The fourth question asked what disclosure would recover, and the answer is most of the window. Disclosing the metric family alone closes 44 per cent, full reconciliation 78 per cent, and consistency disclosure brings the total to 82 per cent, in line with the variance shares. A residual range of 0.21 survives full reconciliation, and 0.16 survives consistency disclosure. Verdict ambiguity falls from all cells to 35, then 27, and finally 12 per cent once only sampling error remains. Both the return and the ceiling matter for policy, and the ceiling is a finding in its own right.

7.2 Theoretical implications

A magnitude changes what the impossibility results can support. A regulator told that fairness metrics differ in principle has no basis for a rule (Kleinberg et al. 2017; Chouldechova 2017). Principle alone does not say whether the gap matters in deployment. A regulator shown that the verdict flips inside the

defensible space for every system tested faces a policy problem instead. That problem concerns disclosure, since the task becomes placing a reported figure within a known range rather than picking the correct metric.

What a reported figure means changes with the magnitude as well. A published score of this kind is a selection from a range whose width the organisation can compute and the reader cannot, rather than an estimate with a confidence interval. An audit that reports one score without the choices behind it certifies a number without certifying the number's governance. Financial reporting reached the same diagnosis and regulated the disclosure rather than the measure (European Securities and Markets Authority 2015; Black et al. 2018).

The variance shares settle whether the three discretion points name one problem three times. They do not. Selection moves the figure most, definition carries a bounded second share, and consistency contributes almost nothing on these panels. These weights hold in a stable order across four unrelated systems. That order tells standard-setters where to aim first. A rule addressed to one point would therefore leave the others untouched, a design constraint that no qualitative taxonomy on its own can supply.

The same weights place the construct against model multiplicity and fairness hacking. Multiplicity research shows that near-optimal models disagree about predictions and group fairness (Marx et al. 2020; Black et al. 2022; Watson-Daniels et al. 2023). Training randomness alone moves group fairness measures (Ganesh et al. 2023; Cooper et al. 2024). The audit space measured here is looser still, and its loosest choice, the metric family, is the least disclosed (Meding and Hagendorff 2024; Simson et al. 2024). Discretion therefore accumulates downstream of model building, where a single figure is chosen for a regulator.

7.3 Implications for practice and policy

For internal audit teams, the findings define a minimum disclosure set that costs nothing beyond writing down choices already made. Every quantified ethics finding should carry four elements: the metric family, its operational definition, the reference population and the group comparison rule. Study 3 prices the step, since those four elements close 78 per cent of the window and cut verdict ambiguity to 35 per cent. Nothing new needs to be measured and no canonical standard is required.

For boards, the magnitude turns record-keeping into an oversight duty. A board handed one reported score cannot judge whether it received the most representative figure or the most favourable one. Nothing in the report reveals the range the figure was drawn from, though the organisation can compute it at will. Requiring the four elements alongside any reported score lets the board challenge the figure it is given. Compliance realism still applies, since ethics work inside firms often cannot change the decisions it reviews (Ali et al. 2023; Rakova et al. 2021).

For regulators, every remedy tested operates on disclosure alone. Each is therefore available under existing rulemaking powers, without first settling which metric is correct. New York City already

mandates published impact ratios, so adding the population, threshold and comparison rule would be a small amendment (Groves et al. 2024; Wright et al. 2024). Canada's assessment collects structured risk responses and no quantified finding at all, a gap in the form itself rather than in compliance with it (Treasury Board of Canada Secretariat 2023).

Absent disclosure, oversight arrives through litigation after deployment. The Dutch SyRI judgment halted a welfare risk-scoring system partly because no published measure allowed the non-discrimination claim to be verified (van Bakkum and Zuiderveen Borgesius 2021). The Apple Card inquiry likewise had to rebuild an evidentiary basis that no prior filing supplied (New York State Department of Financial Services 2021). Both cases show what a regulator inherits when the disclosure layer is empty.

For standard-setters, the gap sits inside standards that already exist. ISO/IEC 42001, IEEE Std 7000 – 2021 and the NIST AI Risk Management Framework all require documented risk decisions (International Organization for Standardization 2023; IEEE 2021; NIST 2023). None requires reconciling a quantified metric to a stated reference or holding it steady across periods. A metric can therefore satisfy every current standard and still sit anywhere in the range measured in Study 1.

Agreement is more tractable here than for the metrics themselves. International standards coordinate best once a technical field agrees on core terms (Cihon 2019). The four elements measured here are a candidate for that agreement, because they are the minimum a reader needs and Study 1 measures what their absence costs. Writing them into existing management-system standards would reach auditors faster than any new regime.

7.4 Limitations and future research

First, the specification space is constructed. Its axes come from published guidance and the regulatory record, so the results describe the discretion available under that guidance rather than the discretion exercised in real engagements. Access to audit working papers would settle the matter and is the most direct extension. Applying the four-fifths threshold uniformly is a device for comparable verdicts rather than a claim of legal exposure. Outside employment selection it has no legal force, and Section 4.9 already restricts the space to its proper scope.

Second, coverage is narrow. One deployed instrument and three benchmark tasks, all binary classification with recorded attributes, leave production audits and generative systems outside the design (Andrus et al. 2021). Production models are tuned more carefully and run on richer features, so nothing here settles whether their windows would be narrower or wider. Three learners on four datasets bound the model-class concern without deciding it. Whether explainability, privacy or robustness metrics show comparable windows is equally open, because the impossibility results behind the fairness case do not transfer automatically.

Third, the corpora are disclosure regimes rather than a sample of audits. New York City's mandated audits, the most direct evidence, cannot yet be collected because no registry exists and posted audits select on willingness to publish (Wright et al. 2024). Manual checking trades one error type for another,

and a document the screen never matched was never read. Study 2's confirmed count of 38 is therefore a lower bound. Registry access would let future work test the disclosure findings where they matter most.

Fourth, Study 3 simulates disclosure rather than observing it. Its estimate assumes accurate and complete reporting, an upper bound on any real regime (Black et al. 2021). The closure figures are also properties of the enumerated space. A regime facing a wider set of choices than this study enumerated would close proportionally less, so the figures bound the mechanism's value under a stated enumeration rather than estimate it in general. A fixed space cannot capture behavioural response, since a salient reference calculation can become a target (Strathern 1997; Espeland and Sauder 2007; Power 1997; Muller 2018). Legal limits on what disclosure duties can reach add a bound no transparency rule dissolves (Wachter et al. 2017; Kroll et al. 2017). Field experiments that assign disclosure rules to live audit engagements are the natural next test.

8 Conclusion

Same system, opposite verdicts describes a measured property of current audit practice. On the evidence assembled here, the same system yields opposite verdicts because the audit, rather than the underlying behaviour, fixes which computed figure becomes the finding. An audit asked for a number, with no standard fixing how the number is computed, can defensibly report almost any figure the scale permits. Even the identity of the disadvantaged group is an artefact of choices no current disclosure records.

Current practice records how each figure was computed and never which figure was chosen. Financial reporting's playbook transfers almost intact: define the measure, reconcile it to a stated reference, apply it consistently and give it no undue prominence (European Securities and Markets Authority 2015). Applied here, reconciliation alone would close 78 per cent of the window today, and consistency disclosure would bring the total to 82 per cent. The residual of 0.16 that survives every disclosure the playbook names is a result rather than a caveat, since it marks what transparency cannot reach under any current standard.

What follows for a regulator, a board or a standard-setter is concrete. Each can already demand the four disclosures that close most of the range, and none needs a canonical metric before acting. Waiting for that canon, which the underlying mathematics may never supply, leaves the range invisible in the meantime. Making the discretion legible is therefore available now. Removing it is not, and the measured residual is the size of that limit.

Declarations

Author Contribution

n/a - sole author

Competing Interests: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Funding: No funding was received for this work.

Data availability: All data analysed in this study are publicly available. The COMPAS records are distributed by ProPublica and the Adult and German Credit datasets by the UCI Machine Learning Repository. The American Community Survey public use microdata are distributed by the United States Census Bureau, the Algorithmic Impact Assessments by the Government of Canada open government portal, and the model cards by Hugging Face. The analysis code that downloads each source, enumerates the specification space, and reproduces every table and figure is provided with this submission.

Ethics approval: Not applicable. The study analyses publicly released, de-identified records and publicly published organisational documents, and involved no human participants.

References

1. Aïvodji U, Arai H, Fortineau O, Gambs S, Hara S, Tapp A (2019) Fairwashing: the risk of rationalization. In: Proceedings of the 36th International Conference on Machine Learning, PMLR 97:161–170. <https://proceedings.mlr.press/v97/aivodji19a.html>
2. Aïvodji U, Arai H, Gambs S, Hara S (2021) Characterizing the risk of fairwashing. In: Advances in Neural Information Processing Systems 34 (NeurIPS 2021), pp 14822–14834. <https://papers.nips.cc/paper/2021/hash/7caf5e22ea3eb8175ab518429c8589a4-Abstract.html>
3. Ali SJ, Christin A, Smart A, Katila R (2023) Walking the walk of AI ethics in technology companies. arXiv:2305.11576. <https://arxiv.org/abs/2305.11576>
4. Andrus M, Spitzer E, Brown J, Xiang A (2021) What we can't measure, we can't understand: challenges to demographic data procurement in the pursuit of fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency:249–260. <https://doi.org/10.1145/3442188.3445888>
5. Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. ProPublica, 23 May 2016. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
6. Bandy J (2021) Problematic Machine Behavior. Proceedings of the ACM on Human-Computer Interaction 5(CSCW1):1–34. <https://doi.org/10.1145/3449148>
7. Bao M, Zhou A, Zottola S, Brubach B, Desmarais S, Horowitz A et al. (2021) It's COMPASlicated: the messy relationship between RAI datasets and algorithmic fairness benchmarks. arXiv:2106.05498. <https://arxiv.org/abs/2106.05498>
8. Barenstein M (2019) ProPublica's COMPAS data revisited. arXiv:1906.04711. <https://arxiv.org/abs/1906.04711>

9. Barocas S, Hardt M, Narayanan A (2023) *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, Cambridge MA
10. Barocas S, Selbst AD (2016) Big data's disparate impact. *Calif Law Rev* 104(3):671–732. <https://doi.org/10.15779/Z38BG31>
11. Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K et al. (2019) AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63(4/5):4:1–4:15. <https://doi.org/10.1147/jrd.2019.2942287>
12. Berg F, Kölbel JF, Rigobon R (2022) Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance* 26(6):1315–1344. <https://doi.org/10.1093/rof/rfac033>
13. Bhattacharya N, Black EL, Christensen TE, Larson CR (2003) Assessing the relative informativeness and permanence of pro forma earnings and GAAP operating earnings. *Journal of Accounting and Economics* 36(1–3):285–319. <https://doi.org/10.1016/j.jacceco.2003.06.001>
14. Binns R (2018) Fairness in machine learning: lessons from political philosophy. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR 81:149–159. <https://proceedings.mlr.press/v81/binns18a.html>
15. Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V et al. (2020) Fairlearn: a toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft Research. <https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/>
16. Birhane A, Steed R, Ojewale V, Vecchione B, Raji ID (2024) AI auditing: The Broken Bus on the Road to AI Accountability. In: *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp 612–643. <https://doi.org/10.1109/satml59370.2024.00037>
17. Black DE, Christensen TE, Ciesielski JT, Whipple BC (2018) Non-GAAP reporting: Evidence from academia and current practice. *Journal of Business Finance & Accounting* 45(3–4):259–294. <https://doi.org/10.1111/jbfa.12298>
18. Black DE, Christensen TE, Ciesielski JT, Whipple BC (2021) Non-GAAP Earnings: A Consistency and Comparability Crisis? *Contemporary Accounting Research* 38(3):1712–1747. <https://doi.org/10.1111/1911-3846.12671>
19. Black E, Raghavan M, Barocas S (2022) Model Multiplicity: Opportunities, Concerns, and Solutions. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp 850–863. <https://doi.org/10.1145/3531146.3533149>
20. Bommasani R, Klyman K, Longpre S, Kapoor S, Maslej N, Xiong B et al. (2023) The Foundation Model Transparency Index. *arXiv:2310.12941*. <https://doi.org/10.48550/arXiv.2310.12941>
21. Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M et al. (2020) Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582(7810):84–88. <https://doi.org/10.1038/s41586-020-2314-9>
22. Bradshaw MT, Sloan RG (2002) GAAP versus The Street: An Empirical Assessment of Two Alternative Definitions of Earnings. *Journal of Accounting Research* 40(1):41–66.

<https://doi.org/10.1111/1475-679x.00038>

23. Breznau N, Rinke EM, Wuttke A, Nguyen HHV, Adem M, Adriaans J et al. (2022) Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences* 119(44).
<https://doi.org/10.1073/pnas.2203150119>
24. Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR 81:77–91. <https://proceedings.mlr.press/v81/buolamwini18a.html>
25. Casper S, Ezell C, Siegmann C, Kolt N, Curtis TL, Bucknall B et al. (2024) Black-Box Access is Insufficient for Rigorous AI Audits. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*:2254–2272. <https://doi.org/10.1145/3630106.3659037>
26. Castelnovo A, Crupi R, Greco G, Regoli D, Penco IG, Cosentini AC (2022) A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12(1):4209.
<https://doi.org/10.1038/s41598-022-07939-1>
27. Chatterji AK, Durand R, Levine DI, Touboul S (2016) Do ratings of firms converge? Implications for managers, investors and strategy researchers. *Strategic Management Journal* 37(8):1597–1614.
<https://doi.org/10.1002/smj.2407>
28. Chouldechova A (2017) Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5(2):153–163. <https://doi.org/10.1089/big.2016.0047>
29. Christensen DM, Serafeim G, Sikochi A (2022) Why is Corporate Virtue in the Eye of The Beholder? The Case of ESG Ratings. *The Accounting Review* 97(1):147–175. <https://doi.org/10.2308/tar-2019-0506>
30. Christensen HB, Hail L, Leuz C (2021) Mandatory CSR and sustainability reporting: economic analysis and literature review. *Review of Accounting Studies* 26(3):1176–1248.
<https://doi.org/10.1007/s11142-021-09609-5>
31. Cihon P (2019) Standards for AI governance: international standards to enable global coordination in AI research and development. Future of Humanity Institute, University of Oxford.
<https://www.governance.ai/research-paper/standards-for-ai-governance-international-standards-to-enable-global-coordination-in-ai-research-development>
32. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B et al. (2024) TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*:e078378. <https://doi.org/10.1136/bmj-2023-078378>
33. Cooper AF, Lee K, Choksi MZ, Barocas S, De Sa C, Grimmelmann J et al. (2024) Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 38(20):22004–22012.
<https://doi.org/10.1609/aaai.v38i20.30203>
34. Corbett-Davies S, Gaebler JD, Nilforoshan H, Shroff R, Goel S (2023) The measure and mismeasure of fairness. *J Mach Learn Res* 24(312):1-117. <https://doi.org/10.48550/arXiv.1808.00023>

35. Costanza-Chock S, Raji ID, Buolamwini J (2022) Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. 2022 ACM Conference on Fairness Accountability and Transparency:1571–1583. <https://doi.org/10.1145/3531146.3533213>
36. Deng WH, Nagireddy M, Lee MSA, Singh J, Wu ZS, Holstein K et al. (2022) Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. 2022 ACM Conference on Fairness Accountability and Transparency:473–484. <https://doi.org/10.1145/3531146.3533113>
37. Dieterich W, Mendoza C, Brennan T (2016) COMPAS risk scales: demonstrating accuracy equity and predictive parity. Northpointe Inc., Traverse City MI.
<https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616>
38. Ding F, Hardt M, Miller J, Schmidt L (2021) Retiring Adult: new datasets for fair machine learning. In: Advances in Neural Information Processing Systems 34 (NeurIPS 2021), pp 6478–6490.
<https://doi.org/10.48550/arXiv.2108.04884>
39. Doyle JT, Jennings JN, Soliman MT (2013) Do managers define non-GAAP earnings to meet or beat analyst forecasts? Journal of Accounting and Economics 56(1):40–56.
<https://doi.org/10.1016/j.jacceco.2013.03.002>
40. Doyle JT, Lundholm RJ, Soliman MT (2003) The Predictive Value of Expenses Excluded from Pro Forma Earnings. Review of Accounting Studies 8(2–3):145–174.
<https://doi.org/10.1023/a:1024472210359>
41. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference:214–226.
<https://doi.org/10.1145/2090236.2090255>
42. Equal Employment Opportunity Commission (1978) Uniform guidelines on employee selection procedures. 29 CFR Part 1607. <https://www.ecfr.gov/current/title-29/subtitle-B/chapter-XIV/part-1607>
43. Espeland WN, Sauder M (2007) Rankings and Reactivity: How Public Measures Recreate Social Worlds. American Journal of Sociology 113(1):1–40. <https://doi.org/10.1086/517897>
44. European Securities and Markets Authority (2015) ESMA guidelines on alternative performance measures. ESMA/2015/1415, Paris. <https://www.esma.europa.eu/document/esma-guidelines-alternative-performance-measures>
45. European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union L, 12 July 2024. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
46. Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and Removing Disparate Impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 259–268. <https://doi.org/10.1145/2783258.2783311>
47. Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. Proceedings of the

- Conference on Fairness, Accountability, and Transparency:329–338.
<https://doi.org/10.1145/3287560.3287589>
48. Ganesh P, Chang H, Strobel M, Shokri R (2023) On The Impact of Machine Learning Randomness on Group Fairness. In: 2023 ACM Conference on Fairness, Accountability, and Transparency, pp 1789–1800. <https://doi.org/10.1145/3593013.3594116>
 49. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H et al. (2021) Datasheets for datasets. *Communications of the ACM* 64(12):86–92. <https://doi.org/10.1145/3458723>
 50. Gelman A, Loken E (2013) The garden of forking paths: why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time. Working paper, Department of Statistics, Columbia University.
<http://www.stat.columbia.edu/~gelman/research/unpublished/forking.pdf>
 51. Goodman EP, Trehu J (2022) AI audit-washing and accountability. German Marshall Fund of the United States, Washington DC. <https://www.gmfus.org/news/ai-audit-washing-and-accountability>
 52. Green B (2022) Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology* 35(4). <https://doi.org/10.1007/s13347-022-00584-6>
 53. Groves L, Metcalf J, Kennedy A, Vecchione B, Strait A (2024) Auditing work: exploring the New York City algorithmic bias audit regime. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), pp 1107–1120.
<https://doi.org/10.1145/3630106.3658959>
 54. Gundersen OE, Kjensmo S (2018) State of the Art: Reproducibility in Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence* 32(1).
<https://doi.org/10.1609/aaai.v32i1.11503>
 55. Hagendorff T (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines* 30(1):99–120. <https://doi.org/10.1007/s11023-020-09517-8>
 56. Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems* 29 (NeurIPS 2016), pp 3315–3323.
<https://doi.org/10.48550/arXiv.1610.02413>
 57. Hellman D (2020) Measuring algorithmic fairness. *Va Law Rev* 106(4):811–866.
<https://virginialawreview.org/articles/measuring-algorithmic-fairness/>
 58. Holstein K, Wortman Vaughan J, Daumé H, Dudik M, Wallach H (2019) Improving fairness in machine learning systems: what do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*:1–16.
<https://doi.org/10.1145/3290605.3300830>
 59. Hutchinson B, Mitchell M (2019) 50 years of test (un)fairness: lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*:49–58.
<https://doi.org/10.1145/3287560.3287600>
 60. IEEE (2021) IEEE Std 7000 – 2021: IEEE standard model process for addressing ethical concerns during system design. Institute of Electrical and Electronics Engineers, New York.

<https://doi.org/10.1109/IEEESTD.2021.9536679>

61. International Organization for Standardization (2023) ISO/IEC 42001:2023 Information technology, artificial intelligence, management system. ISO, Geneva. <https://www.iso.org/standard/42001>
62. Jacobs AZ, Wallach H (2021) Measurement and Fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency:375–385. <https://doi.org/10.1145/3442188.3445901>
63. Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9):389–399. <https://doi.org/10.1038/s42256-019-0088-2>
64. Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent trade-offs in the fair determination of risk scores. In: 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), LIPIcs 67:43:1–43:23. <https://doi.org/10.4230/LIPIcs.ITCS.2017.43>
65. Klevor K, Marquardt CA, McVay SE (2008) SEC Scrutiny and the Evolution of Non-GAAP Reporting. *The Accounting Review* 83(1):157–184. <https://doi.org/10.2308/accr.2008.83.1.157>
66. Kroll JA, Huey J, Barocas S, Felten EW, Reidenberg JR, Robinson DG et al. (2017) Accountable algorithms. *Univ Pa Law Rev* 165(3):633–705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/
67. Lamoreaux PT, Matkaluk L, Sheneman AG (2026) Do Investors Value Auditor Involvement in Non-GAAP Reporting? *Journal of Accounting Research*. <https://doi.org/10.1111/1475-679x.70064>
68. Larson J, Mattu S, Kirchner L, Angwin J (2016) How we analyzed the COMPAS recidivism algorithm. ProPublica, 23 May 2016. <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>
69. Laux J, Wachter S, Mittelstadt B (2024) Three pathways for standardisation and ethical disclosure by default under the EU Artificial Intelligence Act. *Computer Law & Security Review* 53:105957. <https://doi.org/10.1016/j.clsr.2024.105957>
70. Leung E, Veenman D (2018) Non-GAAP Earnings Disclosure in Loss Firms. *Journal of Accounting Research* 56(4):1083–1137. <https://doi.org/10.1111/1475-679x.12216>
71. Liang W, Rajani N, Yang X, Ozoani E, Wu E, Chen Y et al. (2024) Systematic analysis of 32,111 AI model cards characterizes documentation practice in machine learning. *Nature Machine Intelligence* 6(7):744–753. <https://doi.org/10.1038/s42256-024-00857-z>
72. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Chan AW et al. (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine* 26(9):1364–1374. <https://doi.org/10.1038/s41591-020-1034-x>
73. Lougee BA, Marquardt CA (2004) Earnings Informativeness and Strategic Disclosure: An Empirical Examination of “Pro Forma” Earnings. *The Accounting Review* 79(3):769–795. <https://doi.org/10.2308/accr.2004.79.3.769>
74. Lum K, Zhang Y, Bower A (2022) De-biasing “bias” measurement. 2022 ACM Conference on Fairness Accountability and Transparency:379–389. <https://doi.org/10.1145/3531146.3533105>

75. Madaio M, Egede L, Subramonyam H, Wortman Vaughan J, Wallach H (2022) Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction* 6(CSCW1):1–26. <https://doi.org/10.1145/3512899>
76. Marx C, Calmon FP, Ustun B (2020) Predictive multiplicity in classification. In: *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:6765–6774. <https://proceedings.mlr.press/v119/marx20a.html>
77. Meding K, Hagendorff T (2024) Fairness Hacking: The Malicious Practice of Shrouding Unfairness in Algorithms. *Philosophy & Technology* 37(1):4. <https://doi.org/10.1007/s13347-023-00679-8>
78. Metaxa D, Park JS, Robertson RE, Karahalios K, Wilson C, Hancock J et al. (2021) Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human-Computer Interaction* 14(4):272–344. <https://doi.org/10.1561/11000000083>
79. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B et al. (2019) Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*:220–229. <https://doi.org/10.1145/3287560.3287596>
80. Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2021) Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8(1):141–163. <https://doi.org/10.1146/annurev-statistics-042720-125902>
81. Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1(11):501–507. <https://doi.org/10.1038/s42256-019-0114-4>
82. Mökander J, Floridi L (2021) Ethics-Based Auditing to Develop Trustworthy AI. *Minds and Machines* 31(2):323–327. <https://doi.org/10.1007/s11023-021-09557-8>
83. Muller J (2018) *The Tyranny of Metrics*. Princeton University Press. <https://doi.org/10.23943/9781400889433>
84. Munn L (2022) The uselessness of AI ethics. *AI and Ethics* 3(3):869–877. <https://doi.org/10.1007/s43681-022-00209-w>
85. Narayanan A (2018) Translation tutorial: 21 fairness definitions and their politics. Tutorial presented at the Conference on Fairness, Accountability and Transparency, New York, 23 February 2018. <https://fairmlbook.org/tutorial2.html>
86. New York State Department of Financial Services (2021) Report on Apple Card investigation. NYDFS, Albany. https://www.dfs.ny.gov/reports_and_publications/202103_report_apple_card_investigation
87. NIST (2023) Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg. <https://doi.org/10.6028/NIST.AI.100-1>
88. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453. <https://doi.org/10.1126/science.aax2342>

89. Ojewale V, Steed R, Vecchione B, Birhane A, Raji ID (2025) Towards AI accountability infrastructure: gaps and opportunities in AI audit tooling. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3706598.3713301>
90. Pfohl SR, Foryciarz A, Shah NH (2021) An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics* 113:103621. <https://doi.org/10.1016/j.jbi.2020.103621>
91. Pineau J, Vincent-Lamarre P, Sinha K, Lariviere V, Beygelzimer A, Fox E et al. (2021) Improving reproducibility in machine learning research: a report from the NeurIPS 2019 reproducibility program. *J Mach Learn Res* 22(164):1–20. <https://doi.org/10.48550/arXiv.2003.12206>
92. Power M (1997) *The Audit Society: Rituals of Verification*. Oxford University Press, Oxford
93. Raghavan M, Barocas S, Kleinberg J, Levy K (2020) Mitigating bias in algorithmic hiring. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*:469–481. <https://doi.org/10.1145/3351095.3372828>
94. Raji ID, Buolamwini J (2019) Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*:429–435. <https://doi.org/10.1145/3306618.3314244>
95. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B et al. (2020) Closing the AI accountability gap. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*:33–44. <https://doi.org/10.1145/3351095.3372873>
96. Raji ID, Xu P, Honigsberg C, Ho D (2022) Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*:557–571. <https://doi.org/10.1145/3514094.3534181>
97. Rakova B, Yang J, Cramer H, Chowdhury R (2021) Where Responsible AI meets Reality. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW1):1–23. <https://doi.org/10.1145/3449081>
98. Rességuier A, Rodrigues R (2020) AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society* 7(2). <https://doi.org/10.1177/2053951720942541>
99. Rodolfa KT, Salomon E, Haynes L, Mendieta IH, Larson J, Ghani R (2020) Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*:142–153. <https://doi.org/10.1145/3351095.3372863>
100. Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A et al. (2018) Aequitas: a bias and fairness audit toolkit. *arXiv*:1811.05577. <https://doi.org/10.48550/arXiv.1811.05577>
101. Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*:59–68. <https://doi.org/10.1145/3287560.3287598>
102. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E et al. (2018) Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in*

- Methods and Practices in Psychological Science 1(3):337–356.
<https://doi.org/10.1177/2515245917747646>
103. Simmons JP, Nelson LD, Simonsohn U (2011) False-Positive Psychology. *Psychological Science* 22(11):1359–1366. <https://doi.org/10.1177/0956797611417632>
 104. Simnett R, Vanstraelen A, Chua WF (2009) Assurance on Sustainability Reports: An International Comparison. *The Accounting Review* 84(3):937–967. <https://doi.org/10.2308/accr.2009.84.3.937>
 105. Simonsohn U, Simmons JP, Nelson LD (2020) Specification curve analysis. *Nature Human Behaviour* 4(11):1208–1214. <https://doi.org/10.1038/s41562-020-0912-z>
 106. Simson J, Pfisterer F, Kern C (2024) One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp 1305–1320.
<https://doi.org/10.1145/3630106.3658974>
 107. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W (2016) Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* 11(5):702–712.
<https://doi.org/10.1177/1745691616658637>
 108. Strathern M (1997) ‘Improving ratings’: audit in the British University system. *European Review* 5(3):305–321. [https://doi.org/10.1002/\(sici\)1234-981x\(199707\)5:3%3C305::aid-euro184%3E3.0.co;2-4](https://doi.org/10.1002/(sici)1234-981x(199707)5:3%3C305::aid-euro184%3E3.0.co;2-4)
 109. Thomas RL, Uminsky D (2022) Reliance on metrics is a fundamental challenge for AI. *Patterns* 3(5):100476. <https://doi.org/10.1016/j.patter.2022.100476>
 110. Treasury Board of Canada Secretariat (2023) Directive on automated decision-making. Government of Canada, Ottawa. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>
 111. van Bekkum M, Zuiderveen Borgesius F (2021) Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security* 23(4):323–340.
<https://doi.org/10.1177/13882627211031257>
 112. Veale M, Binns R (2017) Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4(2):205395171774353.
<https://doi.org/10.1177/2053951717743530>
 113. Verma S, Rubin J (2018) Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*:1–7. <https://doi.org/10.1145/3194770.3194776>
 114. Wachter S, Mittelstadt B, Floridi L (2017) Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law* 7(2):76–99. <https://doi.org/10.1093/idpl/ix005>
 115. Wachter S, Mittelstadt B, Russell C (2021) Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review* 41:105567.
<https://doi.org/10.1016/j.clsr.2021.105567>
 116. Watson-Daniels J, Parkes DC, Ustun B (2023) Predictive Multiplicity in Probabilistic Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 37(9):10306–10314.

<https://doi.org/10.1609/aaai.v37i9.26227>

117. Wilson C, Ghosh A, Jiang S, Mislove A, Baker L, Szary J et al. (2021) Building and Auditing Fair Algorithms. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency:666–677. <https://doi.org/10.1145/3442188.3445928>

118. Wright L, Muenster RM, Vecchione B, Qu T, Cai PS, Smith A et al. (2024) Null compliance: NYC Local Law 144 and the challenges of algorithm accountability. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), pp 1701–1713. <https://doi.org/10.1145/3630106.3658998>

Figures

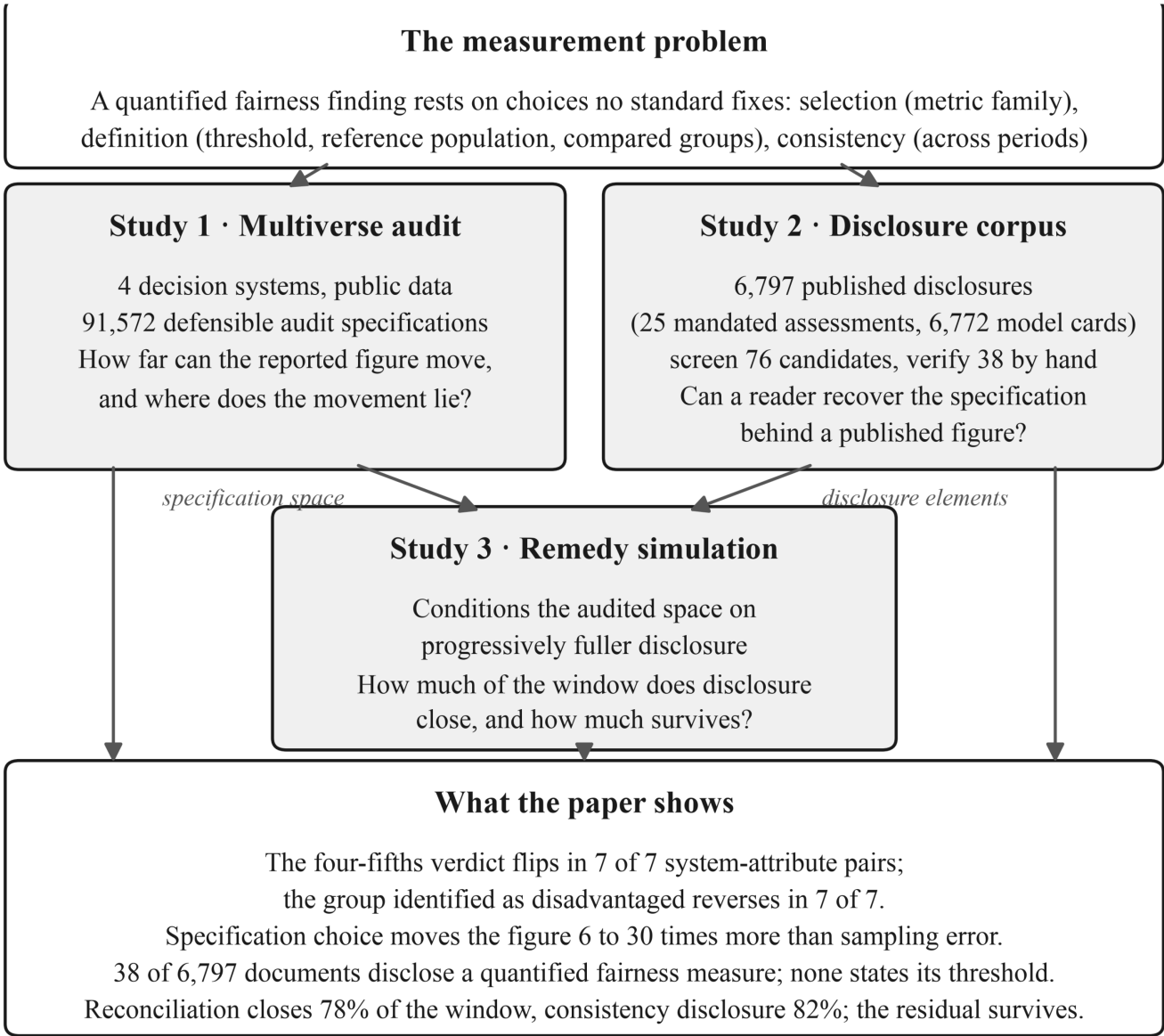


Figure 1

Research design. Study 1 enumerates every defensible audit specification for four decision systems, Study 2 reads every quantified fairness claim in two real disclosure regimes, and Study 3 conditions the audited space on progressively fuller disclosure. All numbers shown are derived in Sections 4 to 6.

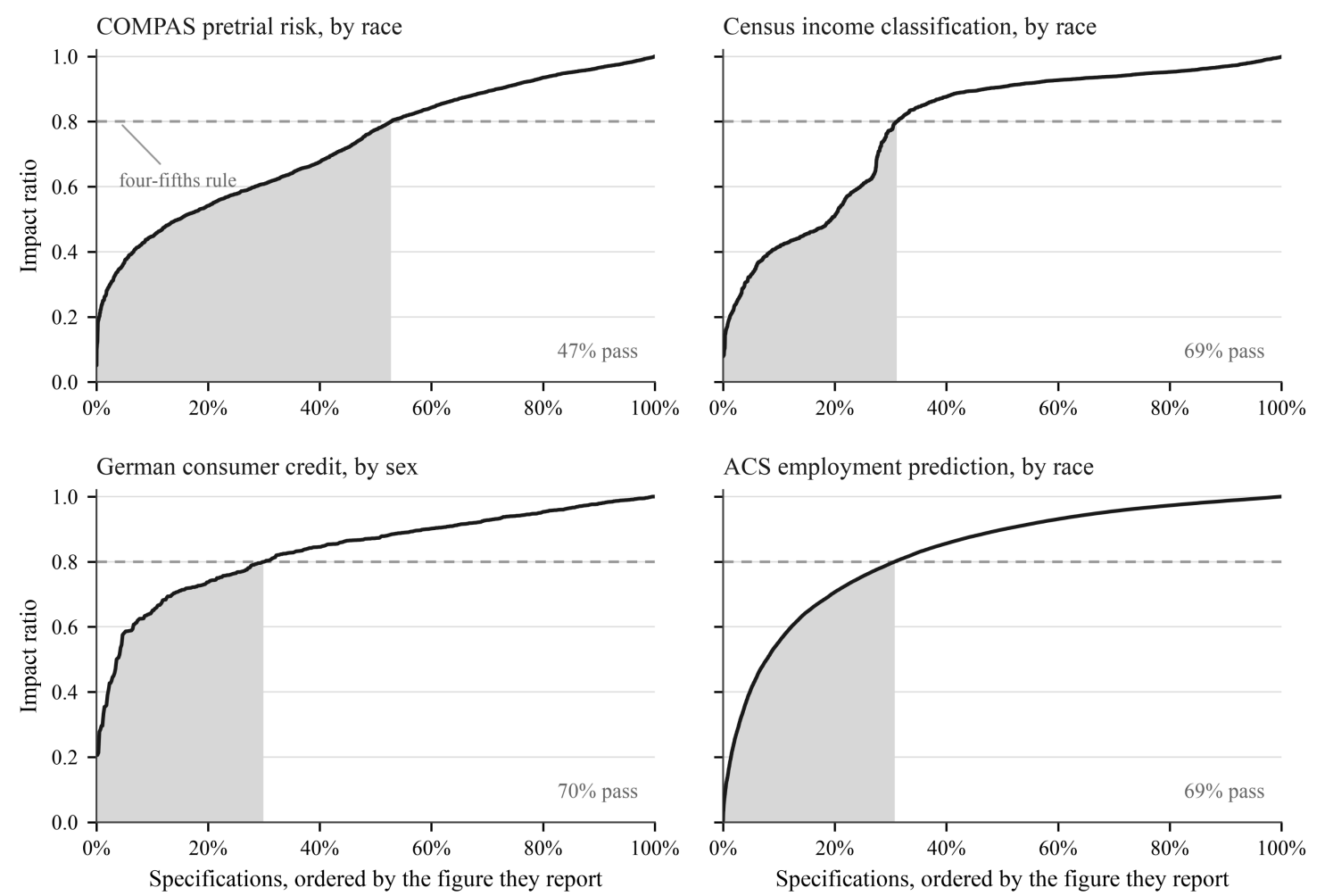


Figure 2

The specification curve for four of the audited systems. Each panel orders every specification in the defensible core by the impact ratio it reports. The dashed line is the four-fifths threshold and the shaded area marks specifications that fail it.

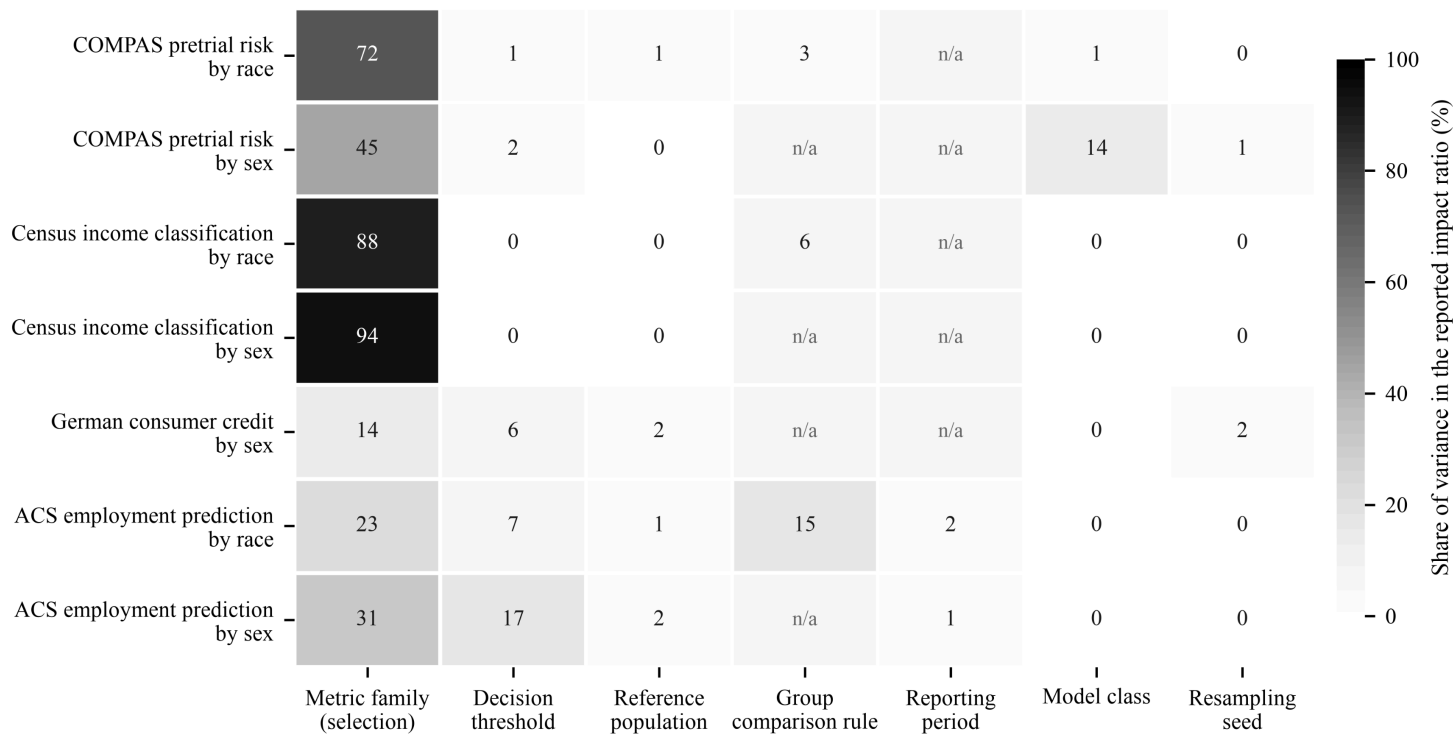


Figure 3

Share of variance in the reported impact ratio attributable to each axis of the specification space, by system and protected attribute. Model class and resampling seed are benchmarks rather than discretion points.

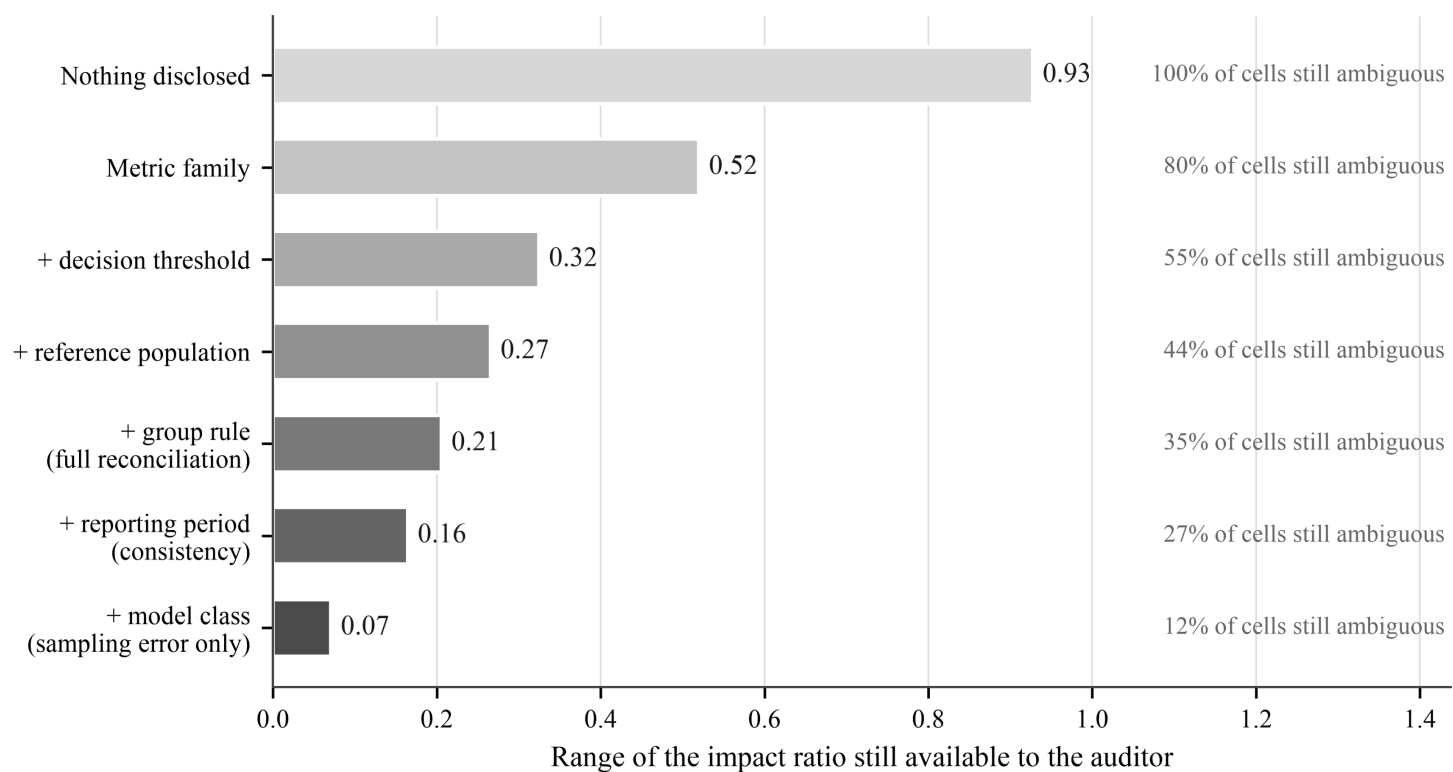


Figure 4

The discretion window under progressive disclosure, pooled across systems. Each bar is the range of impact ratios still consistent with everything the auditor has disclosed to that point.