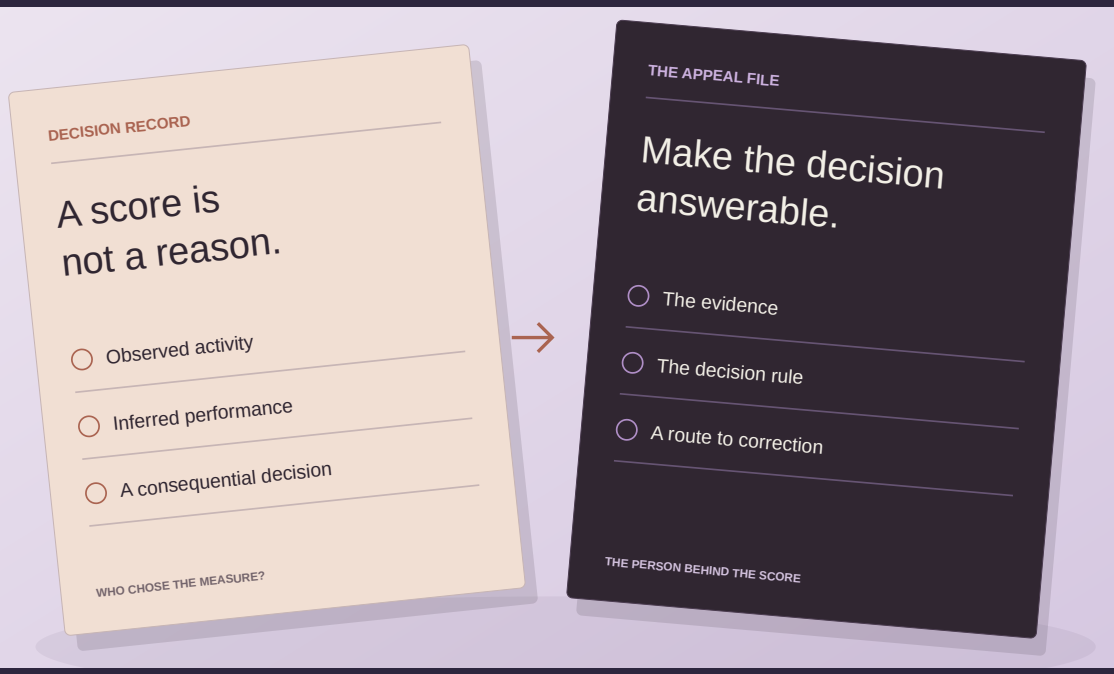
kymata labs

RESEARCH PAPER 04 / AI & WORKPLACE POWER

Your Boss Bought a Score.

When an automated recommendation decides your opportunity, who can change its mind?



Kymata Labs Research
2026-11-16 · Edition 1.0

Independent analysis
kymatalabs.com

Small studio. Open worlds.

THE LONGER VIEW

Executive summary.

A workplace score acquires power when an organization attaches consequences to it. This paper examines the distinction between a prediction, the policy that uses it, and the decision an employee experiences. Inspected Fairlearn and Aequitas documentation supplies the technical foundation: construct validity, distinct fairness metrics, small-group uncertainty, and the separation of predictors from decision thresholds. From those limits, we propose an accountable decision record, a practical appeal chain, and procurement demonstrations that test whether errors can lead to corrected outcomes. A hypothetical promotion example makes the mechanisms concrete without claiming an investigated employer case. The paper addresses the strongest case for automation, privacy tradeoffs and proportionate oversight; it makes no estimate of workplace adoption or legal compliance.

> Mechanism analysis of seven inspected documents from two official software projects, Fairlearn and Aequitas, captured at exact repository revisions on November 16, 2026. No workplace model, personnel dataset or employer case was tested. The promotion scenario and governance framework are explicitly hypothetical and proposed; no adoption, discrimination prevalence or current legal conclusion is inferred.

01 / The recommendation that becomes a fact
02 / What was the score actually measuring?
03 / A number does not choose its own consequences
04 / The fairness badge cannot answer an appeal
05 / The record has to survive the meeting
06 / Who pays for the right to disagree?
07 / Small groups, large uncertainty
08 / The best case for automated recommendations
09 / Buy a process that can change its mind
10 / Evidence and limits
Q&A / Questions worth asking
Sources / References

© 2026 Kymata Labs LLC. All rights reserved.
Read the web edition: kymatalabs.com/research/your-boss-bought-a-score

The recommendation that becomes a fact

Imagine a promotion meeting. This is a hypothetical, not a reported incident. A manager opens a ranked list. Beside each employee is a score, calculated from information the company already collects. The lowest-ranked employee has no invitation to the meeting, no copy of the inputs, and no way to know whether the score helped settle the decision.

Everyone inside the room can plausibly describe the system as advisory. The software supplied a recommendation. The manager made the choice. The company bought a tool. Yet the person outside the room encounters a single, finished consequence: someone else got the opportunity.

This is where the argument about algorithmic management becomes more interesting than a contest between humans and machines. A score can be statistically competent and still enter an indefensible decision process. It can also expose inconsistency that a comfortable human process has concealed. The consequential question is what happens between the number appearing and the opportunity disappearing.

Our thesis is that employers buying consequential scoring systems should also buy the means of contesting their decisions. A usable challenge requires an identifiable decision, a recoverable record, an authorized reviewer, and a remedy that can arrive while it still matters. An explanation of the model, by itself, supplies none of those guarantees.

That is a governance proposal, not a statement of universal legal rights. This paper examines primary technical documentation from Fairlearn and Aequitas, then develops a workplace decision framework from its limits. It does not report a new investigation of an employer or estimate how widely particular systems are deployed. The hypothetical running example makes the mechanisms concrete without borrowing the authority of an incident we have not established.

What was the score actually measuring?

A prediction begins with a choice about the world. “Who will succeed?” must become something observable enough to train or evaluate against. Past promotion? A manager's rating? Sales revenue? Staying with the company? Those outcomes can be relevant. They are not interchangeable definitions of contribution.

Fairlearn's documentation calls attention to construct validity: whether a measurement meaningfully captures the intended concept. Its hiring example explicitly asks whether the label “qualified” is valid before proposing a false-negative rate for candidates screened out. The arithmetic comes after the argument about what the arithmetic represents. [1][2]

Return to the hypothetical promotion system. Suppose its target is previous managers' promotion decisions. A high test score could establish agreement with those decisions, under the evaluation's conditions. It would not establish that the model measures potential independently of the old process. Calling the target “leadership readiness” would not create that missing evidence. The label on the dashboard can be more ambitious than the label in the dataset.

Now suppose the target is next year's performance rating. That seems closer to the intended outcome. It still raises questions about who receives difficult assignments, who gets credit, how ratings vary between managers, and whether the employees whose outcomes were observed resemble those now being ranked. These are questions to investigate, not findings that every performance dataset is defective.

The organization should preserve the translation from aspiration to label. What precisely is predicted, for which population, over what period? Which information is missing? What outcome would count as evidence that the measure was wrong? If the answer to the last question is always “the employee failed to understand the system,” the company has designed a doctrine rather than a testable recommendation.

There is a second translation after prediction. An estimate becomes a ranking; a ranking becomes a shortlist; a shortlist determines who receives attention. Each step can change the practical meaning of the original output. A score's usefulness for inviting further review does not automatically establish its usefulness for denying an opportunity.

Fairlearn's account of allocation harms includes systems that extend or withhold opportunities, resources, or information. That vocabulary is valuable because it asks what the system does to people, rather than whether its designers intended prejudice. [1] An employer's stated intention remains relevant. It cannot substitute for examining the allocation.

A number does not choose its own consequences

There is no natural promotion threshold hidden inside a model. An organization decides how many places exist, what evidence it will consider, how to handle ties, and when an exception is justified. A model may participate in those choices. It does not make them cease to be choices.

Fairlearn's postprocessing documentation offers a concrete example of this separation. Its ThresholdOptimizer applies group-specific thresholds to an existing predictor, using an objective and fairness constraints. The documentation describes possible performance tradeoffs, requires sensitive features at prediction time, and explains that satisfying a constraint can require randomized threshold selection. [6]

This is not a recommendation to use that technique for employment decisions. Its suitability, data use, and legal acceptability require separate assessment. It is evidence that a predictor and the policy converting its output into action are distinct technical objects. The same underlying scores can participate in different decision rules.

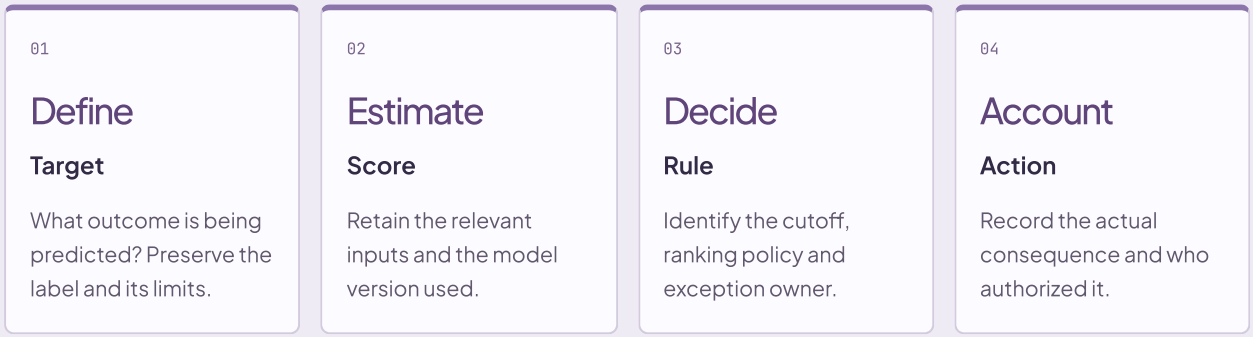
That distinction should survive the procurement process. A vendor can supply evidence about a predictor. The employer still has to justify its use for a particular decision. If a ranking is used only to suggest people for an additional conversation, the consequence differs from automatically excluding everyone below a line. Calling both “decision support” obscures the difference that most matters to the person being scored.

A useful decision record therefore separates four things: the target being estimated, the output produced, the policy applied, and the action taken. It records who owns each transition. This is our proposed architecture, informed by the sources' distinctions rather than a standard validated by them.

FIGURE 01 DECISION ARCHITECTURE · PROPOSED

A score does not choose its consequences.

Keep the prediction separate from the organizational choices that turn it into an opportunity or an exclusion.



An analytical framework, not a measured employer workflow. Threshold methods illustrate the distinction; they are not endorsed for employment use.

Kymata Labs proposed framework, informed by Fairlearn construct-validity and postprocessing documentation. [1] [2] [6]

The organization also needs to inspect indirect decisions. A manager might retain formal authority while seeing only a filtered shortlist. The omitted employees never become visible alternatives. In that hypothetical arrangement, the consequential act occurred before the meeting. Auditing only the manager's final click would miss the filter's role.

The same logic applies to attention. A red flag that triggers an investigation is different from a red flag that triggers a sanction, but an investigation still consumes time and can affect how subsequent evidence is interpreted. Proportional safeguards need a description of the real workflow, including intermediate burdens, not simply a declaration that the final decision is human.

The fairness badge cannot answer an appeal

The strongest technical auditing tools do not promise to settle every question about fairness. Their documentation repeatedly warns against that interpretation.

Fairlearn distinguishes demographic parity, which compares selection rates across groups, from equalized odds, which compares true-positive and false-positive rates. Equal opportunity considers the true-positive rate alone. These criteria address different questions and depend on different assumptions. In particular, error rates depend on what the evaluation treats as a correct label. [3]

For a workplace buyer, the practical implication is that “passed the bias audit” is an incomplete sentence. Which outcome, which groups, which data, which criterion, and what uncertainty? A selection-rate comparison might reveal a disparity deserving investigation. It does not establish that each selected person was evaluated appropriately. An error-rate comparison might be useful when the labels are meaningful. It cannot independently repair a poorly defined target.

Aequitas makes the inputs visible. Its auditing quickstart expects target labels, model outputs, and sensitive attributes, and lets users select metrics such as true-positive rate, false-positive rate, and predicted prevalence. Its documentation describes tools for auditing and mitigation experiments in binary classification. [7] That is useful machinery, with a defined scope. It is not a mechanism for deciding whether an employee's missing assignment should change a promotion judgment.

Fairlearn also explicitly identifies a limitation of group metrics: they do not capture every individual's experience or procedural justice. [1] A company could have acceptable group-level results under its chosen criterion while one employee's record contains a material error. Conversely, correcting a single employee's record would not establish that the overall process is fair. Individual challenge and aggregate audit are complementary jobs.

The distinction matters politically as well as technically. A universal fairness badge can make disagreement appear uninformed: the system was audited, so the complainant must be confused. A properly scoped audit should have the opposite effect. It should make clear which claims were tested, which were not, and where a specific challenge belongs.

We make no claim that a metric threshold proves compliance with current employment law. Fairlearn's own discussion warns against treating a familiar numerical rule as a universal fairness or legal guarantee. [3] Legal obligations vary by setting and require current, appropriate advice. A technical certificate should describe technical evidence without pretending to settle those questions.

The record has to survive the meeting

Consider a second stage of the hypothetical promotion case. The employee learns that a score influenced the shortlist and says a substantial project is missing from their record. Three very different organizations can now produce the same polite response: “A human will review your concern.”

In the first, the reviewer sees only today's dashboard. The source data have changed, the model has been updated, and the original shortlist was not preserved. The reviewer can express sympathy but cannot establish what happened.

In the second, the historical record exists, but the reviewer can only rerun the software. The project is added, the number changes, and the recommendation changes with it. No one is authorized to reconsider the closed decision or determine whether other people were affected by the same omission.

In the third, the reviewer can recover the original inputs and policy, correct the record, assess the consequence, and decide what remedy remains possible. They can also escalate a recurring data failure. Only this arrangement connects an individual challenge to organizational action.

These are constructed scenarios. They isolate the difference between having a contact address and having a working review process. A usable appeal needs all four links: notice, evidence, authority, and remedy. A break anywhere can make the rest ceremonial.

Review question	Minimum record	What it makes possible
What happened?	Decision, date, consequence, decision owner	Identify the action being challenged
What informed it?	Relevant inputs, missingness, provenance, model and policy versions	Reconstruct the recommendation actually used
What can change?	Review grounds, authorized reviewer, response deadline	Correct facts and reconsider their use
What happens next?	Outcome, reasons, remedy, escalation route	Connect review to a practical consequence

The table is a proposed minimum, not a demand to expose every colleague's private information or publish proprietary model weights. Different people need different access. An employee may need their own relevant record and an intelligible account of the criterion. An independent reviewer may need controlled access to additional evidence. A public report can provide aggregate findings without exposing individual personnel files.

Retention also has a cost. A permanent archive of every possible worker attribute would create its own exposure. The proposed record should be limited to the decision's evidentiary needs, protected by access controls, and retained under a defined policy appropriate to its purpose and applicable requirements. “Keep enough to investigate” and “keep everything forever” are very different instructions.

FIGURE 02 REVIEW DESIGN · PROPOSED

An appeal needs a complete chain.

A complaint channel can change a decision only when each link has an owner and usable evidence.

01	02	03	04
Know	**Check**	**Review**	**Act**
Notice	**Evidence**	**Authority**	**Remedy**
Identify the decision, consequence and responsible owner.	Recover the relevant record, criterion and historical version.	Give a reviewer power to correct facts and reconsider use.	Deliver a reasoned outcome while correction can still matter.

Role-based access and defined retention protect personnel information. A group-level metric cannot replace an individual review.

Kymata Labs proposed appeal framework; Fairlearn documents the procedural limits of group fairness metrics. [1] [3]

Who pays for the right to disagree?

An appeal consumes time before it consumes an algorithm. Someone must discover that a decision involved a score, understand what is wrong, gather evidence, submit it, and wait. A process that assumes unlimited time, confidence, literacy, or freedom to challenge a manager can be formally open and practically inaccessible.

That is an analytical concern, not a measured description of all employers. It suggests a concrete procurement question: does the system make a material error easier to find and correct, or merely provide a longer explanation of why the original output appeared?

The distinction between explanation and correction is sharpest when the explanation is entirely accurate. “Your score was lower because this project was absent” explains the computation. It does not justify excluding the project. “You ranked below the cutoff” explains the decision rule. It does not justify the cutoff's use for this opportunity. A coherent reason can still support a challenge.

The employer should fund the machinery needed to resolve that challenge. This includes a reviewer with sufficient time and authority, accessible channels, and a deadline matched to the decision. A reconsideration after a hiring round closes may have less practical value than a brief pause while a material identity error is checked. That does not mean every complaint must freeze every process. It means urgency and reversibility belong in the design.

Review independence should be proportionate too. A routine correction can often be handled near the source. A dispute about whether the organization chose the wrong target, or whether a manager applied the rule inconsistently, may require someone outside that immediate chain. Routing every objection to

the person whose judgment is being challenged risks making review depend on that person's willingness to reconsider.

None of this guarantees a favorable result for the complainant. A defensible process can uphold an unwelcome decision. It should be able to explain what evidence was examined, what changed or did not change, and why. The ability to disagree with a score is not the ability to demand any outcome; it is the ability to bring relevant evidence into an accountable decision.

Small groups, large uncertainty

Aggregate assessment remains essential. Appeals are incomplete evidence: an absence of complaints cannot establish that no one was harmed. Some errors may be invisible to the people affected; some people may choose not to challenge them. These are reasons to inspect the process proactively, not estimates of complaint behavior.

Fairlearn recommends examining relevant groups and intersections. Its documentation warns that some intersections may contain very few records or none, and recommends including counts because random noise affects small groups more strongly. [2][4] A dashboard that displays an apparent disparity without its denominator invites overinterpretation. A dashboard that suppresses every small group can conceal exactly where the evidence is weakest.

The response should preserve uncertainty. Report what was observed, the group counts, the assessment period, and limitations. Fairlearn's confidence-interval documentation describes bootstrap estimation and cautions that more resampling can improve numerical precision without making an unrepresentative underlying sample more accurate. It also warns that overlapping intervals do not establish that quantities are identical. [5]

These observations are especially relevant to an employer with changing roles, uneven team sizes, and infrequent promotions. A large company-wide dataset may still provide little evidence for a particular intersection in a particular job. An overall result cannot fill that gap by assertion.

There is a further judgment in choosing what to control for. Comparing outcomes within job levels can be informative. It can also hide a contested allocation into those levels if that earlier allocation is treated as beyond question. Fairlearn's discussion of control features makes the analogous caution explicit: conditioning on a variable can be problematic when the variable is connected to historical discrimination. [4] The workplace implication is to explain why a comparison is relevant, rather than presenting every statistical adjustment as neutral.

Sometimes the evidence will remain too weak for a confident conclusion. Then the organization should narrow the system's authority, gather better evidence where appropriate, or use an alternative process. Those are our proposed responses. An uncertain audit does not prove discrimination; it also does not justify declaring the decision system cleared.

The best case for automated recommendations

A serious argument must compare the proposed system with an actual alternative. Human-only selection can be inconsistent, poorly recorded, and difficult to challenge. Removing software does not automatically create an accountable workplace.

A well-designed scoring system could make criteria explicit, identify missing data, support repeatable checks, and expose differences that previously went unmeasured. Fairlearn and Aequitas demonstrate that there are inspectable tools for measuring and comparing aspects of a classifier's behavior. [2][7] Their existence does not establish that a particular employer uses them well. It does make blanket claims that algorithmic decisions are inherently unauditable too crude.

The strongest buyer's case is therefore comparative: this system, used in this workflow, improves consequential decisions relative to the available alternative. Test that proposition against relevant outcomes, error burdens, review effort, and the ability to correct mistakes. Do not compare a documented model with an imaginary human process that has no prejudice, no delay, and no missing records.

Nor should oversight become a ritual designed to prevent any useful change. Safeguards should increase with the seriousness and irreversibility of the consequence. A tool suggesting optional training material needs different controls from a system filtering people out of advancement. Both deserve scrutiny, but identical review machinery would waste resources and obscure priorities.

The practical challenge is institutional incentives. If managers are rewarded solely for clearing a queue, requiring a reason for an override may make accepting the score easier than questioning it. If reviewers are penalized for reversals, a nominal appeal channel may become a defense of the original decision. These are plausible mechanisms to test locally, not observed effects established by our source set.

That is why operational ownership matters. A company should name a person accountable for the whole decision process, with access to information from the supplier, authority over the internal workflow, and responsibility for unresolved challenges. Dividing engineering, procurement, personnel, and compliance tasks is sensible. Dividing them so thoroughly that no one owns the consequence is not.

Buy a process that can change its mind

A procurement review can begin with three demonstrations. First, trace one representative recommendation from its target definition through the information used to the eventual action. Second, introduce a known material error and show how it is discovered, corrected, and assessed for consequence. Third, show how a repeated failure leads to a policy change or a pause in use. Use synthetic or appropriately authorized test records; do not manufacture a real personnel dispute for a demo.

FIGURE 03 PROCUREMENT DEMONSTRATIONS · PROPOSED

Test whether the system can change its mind.

Use synthetic or authorized records to test accountability before attaching consequential decisions to a score.

01

Recover

Trace a decision

Follow target → input → score → rule → action, with owners.

02

Reconsider

Correct an error

Repair a known omission and assess the original consequence.

03

Escalate

Fix the pattern

Find related failures; revise policy or pause use when needed.

These are unvalidated governance tests, not completed workplace experiments or a legal compliance certificate.

Original Kymata Labs governance proposal; no empirical values are depicted.

The second demonstration should test more than whether a form accepts a complaint. Can the reviewer retrieve the original version? Can the decision be reconsidered without silently overwriting history? Are downstream systems told about the correction? If the answer cannot be reconstructed without the supplier's help, is that help actually available under the service arrangement?

The third demonstration distinguishes correction from learning. Fixing one record may leave the same defective field in every other record. A responsible process asks whether the failure is isolated, whether similarly affected decisions need review, and whether continued use is justified. The answer may be a small data fix, a revised criterion, additional evidence, or withdrawing the system from that decision.

Measure the review process without turning it into another misleading score. Time to resolution matters, but a fast refusal can be useless. Reversal rate matters, but a low rate may reflect sound initial decisions or inaccessible review. Track reasons, unresolved cases, and whether remedies were delivered alongside simple totals. Interpret those observations against the decision context rather than treating one number as proof of accountability.

The proposed standard is demanding in one respect: the company must be able to act on what it learns. It is modest in another: it requires neither a perfectly fair model nor universal agreement about merit. It requires a route by which a consequential mistake can become a corrected decision.

Evidence and limits

This paper is a mechanism analysis using seven inspected documents from two primary software projects, not seven independent empirical studies. Fairlearn supplies definitions, assessment methods, statistical cautions, and a thresholding implementation description. Aequitas supplies an inspectable

auditing interface and its stated scope. Exact source files, repository revisions, and hashes were retained on November 16, 2026. [1][2][3][4][5][6][7]

The sources' own citations to academic work were not independently reviewed for this paper and are not presented as separately inspected evidence. No workplace model was executed, no personnel dataset was analyzed, and no estimate of present adoption, discrimination prevalence, wage effects, or legal compliance is made. The promotion story is hypothetical throughout. The decision record, appeal design, and procurement demonstrations are Kymata's proposals; their effectiveness has not been validated as a combined intervention.

Within that boundary, the conclusion is consequential. A score becomes powerful because an organization attaches consequences to it. Accuracy evidence belongs in that organization. So do ownership, access to relevant facts, and the capacity to revise an outcome.

Your boss may have bought a score. The company still owes an account of the decision it chose to make.

Questions worth asking.

Does an employer need AI to make an unfair decision?

No. Human-only processes can also be inconsistent, poorly recorded and difficult to challenge. The appropriate comparison is the proposed workflow against the actual alternative, including outcomes, errors and the ability to correct mistakes. This article does not estimate whether automated or human decisions are generally fairer.

Does passing a bias audit mean the score is fair?

Only a claim within the audit's stated scope is supported. Selection-rate and error-rate metrics answer different questions and depend on groups, data, labels and uncertainty. Group-level results do not establish that every individual record is correct or that a meaningful appeal exists. A numerical threshold is not a universal legal guarantee. [1][3]

What should someone be able to challenge about an AI employee score?

Our proposed process covers material input errors, whether the criterion fits the decision, how the policy was applied, and the consequence that followed. A reviewer needs relevant evidence and authority to act. These are governance recommendations, not a statement of rights available in every jurisdiction.

Would meaningful transparency expose coworkers' private information?

It need not. Different roles can receive different access: the employee gets relevant information about their own case, an authorized reviewer can inspect necessary additional evidence, and public reporting can remain aggregate. The paper proposes limited retention and access controls rather than indiscriminate disclosure.

What should a buyer ask the vendor to demonstrate?

Trace a representative recommendation into its action, correct a known material error in an authorized test record, and show how a recurring failure changes the process. Verify historical versions, reviewer authority, downstream correction and a meaningful remedy. These proposed demonstrations test the operating process; they are not validated proof of fairness or legal compliance.

SOURCES, NOT SCENERY

References.

1. **Fairness in Machine Learning**
   Fairlearn contributors · Undated living documentation; captured 2026–11–16 · Primary software documentation

   Official repository commit f75ec4c4c45c42836516f35f5acf29272c635b09. Definitions of allocation harms and construct validity; inspected lines 7–155 and 275–400, particularly the formalism-trap warning that group metrics omit individual experiences and procedural justice. Conceptual documentation, not workplace prevalence or a legal standard. Underlying cited academic papers were not independently inspected.

   https://github.com/fairlearn/fairlearn/blob/f75ec4c4c45c42836516f35f5acf29272c635b09/docs/user_guide/fairness_in_machine_learning.rst

2. **Performing a Fairness Assessment**
   Fairlearn contributors · Undated living documentation; captured 2026–11–16 · Primary software documentation

   Official repository commit f75ec4c4c45c42836516f35f5acf29272c635b09. Inspected lines 7–175: harms-first assessment, relevant groups and intersections, hiring false-negative example conditioned on construct validity of qualified, proxy validity, disaggregated metrics and group counts. No employee data or model was evaluated for this publication.

   https://github.com/fairlearn/fairlearn/blob/f75ec4c4c45c42836516f35f5acf29272c635b09/docs/user_guide/assessment/perform_fairness_assessment.rst

3. **Common fairness metrics**
   Fairlearn contributors · Undated living documentation; captured 2026–11–16 · Primary software documentation

   Official repository commit f75ec4c4c45c42836516f35f5acf29272c635b09. Inspected demographic parity definition, equalized odds and equal opportunity sections, target-label caveat, and warning against universal use of the four-fifths threshold (lines 33–61, 201–310, 314–343). Used for metric distinctions and limitations, not legal advice. Does not establish any employer passed or failed an audit.

   https://github.com/fairlearn/fairlearn/blob/f75ec4c4c45c42836516f35f5acf29272c635b09/docs/user_guide/assessment/common_fairness_metrics.rst

4. **Intersecting Groups**
   Fairlearn contributors · Undated living documentation; captured 2026–11–16 · Primary software documentation

   Official repository commit f75ec4c4c45c42836516f35f5acf29272c635b09. Inspected lines 7–150: small or empty intersections, counts, sensitive versus control features, and warning that control features can relate to historic discrimination. Workplace examples in our paper are analytical applications, not cases from this source.

   https://github.com/fairlearn/fairlearn/blob/f75ec4c4c45c42836516f35f5acf29272c635b09/docs/user_guide/assessment/intersecting_groups.rst

5. **Confidence Interval Estimation**
   Fairlearn contributors · Undated living documentation; captured 2026–11–16 · Primary software documentation

   Official repository commit f75ec4c4c45c42836516f35f5acf29272c635b09. Inspected lines 7–96: bootstrap resampling, reliance on sample representativeness, more precision not necessarily more accuracy, overlapping intervals not evidence of equality. No interval estimated by Kymata; no numeric example from this source is presented as a workplace result.

   https://github.com/fairlearn/fairlearn/blob/f75ec4c4c45c42836516f35f5acf29272c635b09/docs/user_guide/assessment/confidence_interval_estimation.rst

6. **Postprocessing: Threshold optimizer**
   Fairlearn contributors · Undated living documentation; captured 2026–11–16 · Primary software documentation

   Official repository commit f75ec4c4c45c42836516f35f5acf29272c635b09. Inspected lines 7–45: group-specific thresholds, objective and constraint choices, possible performance tradeoffs, sensitive features required at prediction time, randomization between thresholds. Technical mechanism only; no claim of guaranteed population fairness, lawful employment suitability, or recommended use.

   https://github.com/fairlearn/fairlearn/blob/f75ec4c4c45c42836516f35f5acf29272c635b09/docs/user_guide/mitigation/postprocessing.rst

7. **Aequitas: Bias Auditing & Correction Toolkit — auditing interface**
   DSSG / Aequitas contributors · Undated living documentation; captured 2026–11–16 · Primary software documentation

   Official repository commit b35e9d5fafc3844c6761f1f5b39f5f7669391425. Inspected README lines 7–12 and 44–85: stated binary-classification scope, auditing inputs label_value, score and sensitive attributes, selected metrics tpr/fpr/pprev. Source examples are not evidence of a workplace deployment or audited employer; package was not installed or executed.

   https://github.com/dssg/aequitas/blob/b35e9d5fafc3844c6761f1f5b39f5f7669391425/README.md