F
FAIR LABS
FAIR ARTIFICIAL INTELLIGENCE RESEARCH LABS

TOOLKIT ◆ FAIRNESS & BIAS ◆ 2026

Auditing Algorithmic Bias in Practice

A practical toolkit for identifying and mitigating bias in deployed AI systems—scoping, measurement, testing, remediation, and governance.

PROGRAM	DOCUMENT TYPE	AUDIENCE	PUBLISHED
Fairness & Bias	Toolkit	Practitioners & Auditors	2026

FAIR Labs · A 501(c)(3) nonprofit research institute · Washington, DC

CONTENTS

Table of Contents

EXECUTIVE SUMMARY

The bottom line for auditors

IN ONE PARAGRAPH

You cannot audit bias by running one metric and comparing it to a threshold. A credible audit fixes a scope, chooses a fairness definition that fits the decision and its harms, checks whether the data can even support the measurement, computes a small battery of complementary tests rather than a single number, and reports disparities with the uncertainty and the assumptions attached. Where it finds harm, it reaches first for the cheapest durable fix—often a change to the data, the threshold, or the human workflow, not a new model—and it writes the whole thing down so an independent party can check it and repeat it later. This toolkit supplies each step as a ready-to-run instrument.

3

families of fairness definition—group, individual, and causal—that an audit must choose among

7

stages in the repeatable audit protocol, from scoping to re-audit

3

points of intervention for remediation: pre-processing, in-processing, and post-processing

What we recommend

1. **Choose the fairness definition before you measure.**
Group parity, calibration, and individual fairness cannot generally be satisfied at once. Decide which harm the decision most needs to avoid, pick the metric that speaks to it, and record why—so the choice is a governed judgment rather than a convenient default.

2. **Audit the data before the model.**
Proxies for protected attributes, undersampled groups, and biased labels manufacture disparity that no downstream metric can distinguish from the model's own behavior. Interrogate provenance and representativeness first; a clean metric on dirty data is a false negative.

3 **Report a battery, not a verdict.**
Present error-rate parity, calibration, and disparate-impact results side by side with confidence intervals and subgroup breakdowns. A single passing number hides the tradeoff it silently chose and the small subgroup it silently failed.

4 **Make the audit independent and repeatable.**
An audit graded by the team that built the system, or one that cannot be rerun on the next model version, is advocacy, not assurance. Fund documentation, thresholds, cadence, and independent re-auditing as standing commitments.

The remainder of this toolkit develops each recommendation into operational detail, with checklists, metric tables, and a concrete step-by-step protocol an audit team can adopt directly.

SECTION I

What a Bias Audit Can and Cannot Accomplish

The word *audit* carries a promise it cannot always keep. In finance an audit ends in an opinion—the accounts are fairly stated, or they are not—and the reader is entitled to lean on that opinion. A bias audit is different in kind. It does not certify that a system is fair, because fairness is not a single measurable quantity waiting to be checked off. It produces something more modest and more useful: evidence, gathered under stated assumptions, about how a system distributes its errors and its outcomes across the people it touches, together with a judgment about whether those distributions are acceptable for the decision at hand. Confusing the two—treating an audit as a fairness certificate—is the first and most consequential mistake a team can make.

An audit is a flashlight, not a seal of approval. Pointed carefully, it illuminates disparities that were invisible in aggregate performance. Pointed carelessly, it produces a reassuring number that launders a system's harms into compliance. The difference lies almost entirely in the rigor of the scoping, the honesty of the data assessment, and the willingness to report what the audit could not see. This section sets expectations before the toolkit hands you the instruments, because instruments used against the wrong expectations do damage.

What an audit can do

A well-run bias audit accomplishes several concrete things. It **quantifies disparity**: it replaces the intuition that a system "seems unfair" with measured differences in error rates, selection rates, or calibration across defined groups. It **localizes harm**: it moves from a global average that looks acceptable to the specific subgroup, region, or interaction where the system underperforms. It **tests hypotheses about cause**: it distinguishes, as far as the data allow, disparity that originates in the training labels from disparity introduced by the model or by the deployment workflow. And it **creates a record**: it produces documentation that a regulator, an oversight body, or a future auditor can examine and build upon. Each of these is valuable precisely because none of them requires the audit to pronounce a final verdict.

What an audit cannot do

The limits are as important as the capabilities, and they are routinely oversold. An audit cannot **prove a system is fair** in any absolute sense; at best it can show that under a chosen definition, on the population it examined, within the precision its sample allowed, the measured disparities fell within a stated bound. It cannot **settle the normative question** of which fairness definition is right for a given decision—that is a value judgment the audit can inform but not resolve. It cannot **see groups it did not measure**: harm concentrated in an intersectional subgroup the audit never disaggregated will pass through undetected. And it cannot **guarantee the future**: an audit describes the system as it

was at a moment, on data as it was then, and says nothing durable about a model that will be retrained, a population that will shift, or a workflow that will drift.

A WORKING DEFINITION

Throughout this toolkit, a **bias audit** is a structured investigation of whether a deployed AI system distributes benefits and harms inequitably across groups or individuals, judged against an explicitly chosen fairness definition and the harms the deployment makes possible. The unit audited is the *system in its context*—model, data, workflow, and the humans around it—never the model file in isolation.

Bias is sociotechnical, and so is its audit

The most common technical error in bias auditing is to treat the model as the whole system. Consider a risk score that a caseworker consults before making a decision. The score may be well calibrated across groups on its own terms, and yet the deployment can still produce disparate outcomes because caseworkers trust the score more for some applicants than others, or because the threshold at which the score triggers an intervention was set with one population in mind. The disparity is real, it is measurable in outcomes, and it is invisible to any test that examines only the model's raw predictions. An audit scoped to the model alone would return a clean bill of health for a system that is harming people. This is why every instrument in this toolkit is scoped to the deployment, and why the protocol in Section VI insists on measuring outcomes as experienced, not just predictions as emitted.

> *A bias audit does not tell you a system is fair. It tells you where, for whom, and under what assumptions it is not—which is the only honest place to start.*
>
> — The governing premise of this toolkit

Who the audit serves

An audit has more than one audience, and its usefulness depends on serving each. The **builders** need findings precise enough to act on—not "the model is biased" but "false negative rates differ by a measured margin for this subgroup, concentrated in this region, and here is the most likely cause." **Decision-makers** need the disparities translated into the language of harm and risk, so they can weigh them against the system's benefits. The **affected public** needs a record that someone looked, honestly, and can be held to what they found. And an **independent reviewer** needs enough methodological detail to rerun the audit and reach the same conclusions. A report that speaks only to engineers, in the private dialect of a fairness library, fails three of its four audiences.

With expectations set, we turn to the hardest conceptual problem in the field: what fairness even means, and why you cannot have all of it at once.

SECTION II

Defining Fairness & the Impossibility Results

There is no single definition of fairness, and the leading definitions are provably incompatible. An audit that does not choose deliberately among them has chosen by default, usually badly.

Practitioners often arrive expecting that "fairness" names one thing that a good enough model can achieve. It does not. Fairness is a family of formal criteria, each capturing a different intuition about what equal treatment requires, and the mathematics is unambiguous that you cannot satisfy the major group criteria simultaneously except in degenerate cases. This is not a temporary limitation of current methods; it is a structural feature of the problem. The audit's first intellectual task is therefore to choose which fairness you are auditing for, and to justify that choice against the decision and its harms.

Group fairness: parity across protected groups

Group notions of fairness compare aggregate statistics across protected groups—defined by race, sex, age, disability, or another attribute the decision must not discriminate on. The three that recur most often in practice are worth naming precisely, because they conflict. **Demographic parity** (also called statistical parity) asks that the selection rate—the fraction of people who receive the positive outcome—be equal across groups, regardless of the underlying label. **Equalized odds** asks that the error rates be equal: the true positive rate and the false positive rate should match across groups, so the system is equally accurate for everyone. **Calibration** (or sufficiency) asks that a given score mean the same thing for everyone: among all people assigned a risk of, say, 0.7, the same fraction should actually experience the outcome, whatever their group.

THREE GROUP-FAIRNESS CRITERIA AND WHAT EACH PROTECTS AGAINST

Criterion	Requires equal...	Protects against	Blind spot
Demographic parity	Selection (positive-outcome) rate across groups	Systematic under-selection of a group	Ignores real differences in the base rate; can force unqualified selections
Equalized odds	True-positive and false-positive rates across groups	A group bearing more of the system's mistakes	Can be unachievable jointly with calibration when base rates differ
Calibration / sufficiency	Meaning of a score across groups	Scores that systematically over- or under-state risk for a group	Compatible with very different error rates across groups

Each criterion encodes a defensible idea of fairness, and each fails in a way the others catch. A system can be perfectly calibrated—its scores mean the same thing for everyone—while still assigning far more false positives to one group than another. A system can achieve demographic parity while denying qualified people in one group to make the numbers match. There is no criterion that is simply correct; there is only the criterion that best fits the specific decision, its harms, and the law that governs it.

The impossibility results

The reason these criteria cannot be reconciled is not sloppiness in their formulation; it is a theorem. When the underlying base rate of the outcome genuinely differs between two groups—when the event the model predicts is actually more common in one group than another—it is mathematically impossible for a single imperfect classifier to be calibrated *and* to equalize false positive and false negative rates at the same time. Satisfy calibration and the error rates must diverge; force the error rates to match and calibration breaks. This result, established in the work surrounding recidivism-risk scoring, is the central fact the field organizes itself around. A closely related result shows the same tension between demographic parity and accuracy when base rates differ.

THE TRADEOFF YOU CANNOT ESCAPE

When base rates differ across groups, calibration and equalized error rates cannot both hold. This is a theorem, not a tuning problem. An audit's job is not to make the tradeoff disappear—it cannot—but to name which side of it the system falls on, decide whether that is the right side for this decision, and record the reasoning so the choice is accountable.

The practical consequence is liberating rather than paralyzing. Because you cannot have everything, you are freed from the fantasy of a universally fair model and required instead to reason about harm. In a setting where a false positive is catastrophic for the individual—wrongful denial of a benefit, erroneous flagging for investigation—minimizing disparity in false positive rates may matter most. In a setting where the score feeds a human who needs to trust that a 0.7 means the same thing for everyone, calibration may take priority. The theorem does not tell you which to choose; it tells you that pretending you don't have to choose is incoherent.

Individual and causal fairness

Group fairness is not the only frame. **Individual fairness** asks that similar individuals receive similar treatment—that the system not draw arbitrary distinctions between two people who are alike in every way relevant to the decision. It is intuitively compelling and practically hard, because it requires a defensible notion of "similar for this purpose," and that similarity metric can itself smuggle in the very bias one is trying to remove. **Causal fairness** goes further and asks whether the protected attribute exerts an inappropriate causal influence on the outcome—whether, had a person been of a different

group but otherwise identical, the decision would have changed. Causal approaches are conceptually powerful because they distinguish legitimate from illegitimate pathways of influence, but they demand a causal model of the world that is often contested and rarely fully identifiable from observational data.

A **Use group fairness for population-level accountability.**
When the question is whether a system systematically disadvantages a protected group, group metrics are the right tool and often the ones the law speaks in.

B **Use individual fairness to catch arbitrary distinctions.**
When two comparable applicants receive very different treatment, an individual-fairness lens surfaces it even where group statistics look balanced.

C **Use causal reasoning to interrogate mechanism.**
When you need to know whether a protected attribute is doing illegitimate work—directly or through a proxy—a causal frame forces the question into the open, provided you can defend the causal model.

From definitions to a chosen bar

The audit does not need to adjudicate the philosophy of fairness. It needs to convert the choice into a concrete, recorded commitment: for *this* decision, we prioritize *this* criterion, we will tolerate disparity up to *this* bound, and we chose this way for *these* reasons of harm and law. A definition chosen in the open, and written down, is defensible even when reasonable people would have chosen differently. A definition chosen silently—by using whatever a library reports by default—is indefensible even when it happens to be right.

SECTION III

Scoping an Audit

Most failed audits fail at the beginning, not the end. A team that has not drawn the boundary of the system, named the groups it must protect, and enumerated the harms it is looking for will produce metrics that answer questions no one asked. Scoping is the unglamorous work that determines whether everything downstream is meaningful. It is also cheap: a scoping decision that takes an afternoon can save an audit that would otherwise measure the wrong thing precisely.

Scoping answers three questions in order. Where does the system begin and end? Which attributes and groups are we protecting, and how will we even observe them? And what specific harms, to whom, are we trying to detect? Each question has traps, and each trap has a standard countermeasure. This section walks them.

Drawing the system boundary

The system boundary defines what is inside the audit and what is treated as fixed context. Draw it too tightly—around the model's raw predictions—and you miss the threshold, the workflow, and the human decisions where much real-world disparity is produced. Draw it too loosely—around an entire agency's operations—and the audit becomes unrunnable. The right boundary is the smallest one that contains every component whose behavior can produce or amplify the disparity you care about: the input features and the pipeline that produces them, the model, the decision threshold or policy applied to its output, the human step that consumes it, and the outcome the affected person ultimately experiences.

SCOPING CHECKLIST — RESOLVE EACH ITEM BEFORE MEASUREMENT BEGINS

Scoping question	What to determine	Common failure if skipped
System boundary	Every component from input pipeline to experienced outcome that can produce disparity	Auditing raw predictions and missing threshold or workflow bias
Decision & stakes	What the system decides or influences, and how reversible its errors are	Applying a low-stakes bar to a high-stakes adjudication
Protected attributes	Which groups the law and mission require protecting; how each is observed or inferred	Protecting only attributes that happen to be in the dataset
Reference population	The real population the system serves, against which representativeness is judged	Measuring fairness on a sample that omits the affected groups
Harms of concern	The concrete harms—allocative, quality-of-service, dignitary—and who bears them	Chasing metric parity while the actual harm goes unnamed
Fairness definition	The criterion prioritized and the tolerated disparity bound, with rationale	Letting a library default silently pick the normative standard
Data availability	Whether labels, group membership, and outcomes can be observed at needed granularity	Designing tests the data cannot support, then reporting them anyway

Protected attributes and the problem of observing them

An audit of group fairness requires knowing who belongs to which group—and that information is often precisely what the organization, for good legal and ethical reasons, did not collect. This creates a genuine dilemma. Without group labels, disparity across groups cannot be measured directly. With them, the organization holds sensitive data that carries its own risks. There is no costless resolution, but there are disciplined options: collecting attributes with consent and strict access controls for the narrow purpose of auditing; using validated statistical inference of group membership while treating the inferred labels as uncertain and reporting the added uncertainty; or partnering with an independent auditor who can hold the sensitive data under stronger protections than the deploying team. What is not acceptable is to declare an audit impossible because the attribute is inconvenient to observe, or to infer group membership with a crude proxy and then treat the result as ground truth.

Intersectionality complicates this further and cannot be waved away. Harm often concentrates not in a single protected group but at the intersection of several—an effect invisible to any audit that examines race and sex separately but never together. The scoping decision must state which intersections will be examined, constrained by the reality that sample sizes shrink at every intersection and eventually become too small to measure reliably. Naming that limit honestly in the scope is far better than discovering it as a silent gap in the findings.

SCOPING TRAP: THE CONVENIENT ATTRIBUTE

Teams tend to audit the protected attributes that happen to be in the dataset and quietly ignore the ones that are not. But the decision about which groups deserve protection is a normative and legal question, not a data-availability question. If a legally protected group cannot be observed, that is a finding to report and a gap to close—not a reason to leave the group out of scope.

Enumerating the harms of concern

Fairness metrics are proxies for harms, and an audit that loses sight of the harm optimizes the proxy. It helps to enumerate harms explicitly and by type. **Allocative harms** occur when the system distributes a resource or opportunity unequally—loans, jobs, benefits, bail. **Quality-of-service harms** occur when the system simply works worse for some groups—speech recognition that fails on some accents, image analysis that fails on some skin tones. **Dignitary and representational harms** occur when the system demeans or stereotypes, independent of any resource it allocates. Different harms call for different metrics and different remedies, and a single audit often needs to address more than one. Writing the harms down, and mapping each to the metric that will detect it, keeps the audit anchored to what actually matters to the people it is meant to protect.

> *Scope is where an audit decides what it is willing to see. Everything it leaves outside the boundary, it promises not to find.*
>
> — On the discipline of scoping

Writing the scope down

The deliverable of scoping is a short written scope statement: the system boundary, the protected attributes and intersections in scope, the reference population, the enumerated harms, the chosen fairness definition and disparity bound, and an honest note on what the available data will and will not support. This document is not bureaucratic overhead; it is the audit's constitution. When a finding is later disputed, the scope statement is what distinguishes a deliberate boundary from an accidental blind spot, and it is the first thing an independent re-auditor will ask to see.

SECTION IV

Data & Measurement

Most algorithmic bias is inherited, not invented. It enters through the data—through what was measured, who was counted, and how the labels were made—long before the model is trained. An audit that starts at the model starts too late.

A model is a compression of its training data. If the data encode a historical pattern of unequal treatment, the model will reproduce that pattern faithfully and call it accuracy. If the data undercount a group, the model will be worse for that group and the aggregate metric will hide it. If the labels the model learned from were themselves the product of biased human decisions, the model will learn to replicate the bias with mechanical consistency. This section is a guide to interrogating the data before trusting any downstream number, because the cleanest metric in the world computed on compromised data is a confident false negative.

Proxies: bias that survives the removal of the protected attribute

The most persistent misconception in practice is that removing the protected attribute from the model's inputs removes the bias. It does not, because the world is dense with proxies. A postal code correlates with race; a first name correlates with sex and ethnicity; a gap in employment history correlates with disability, caregiving, or incarceration. When the protected attribute is dropped but its proxies remain, the model reconstructs the attribute from the proxies and discriminates just as effectively, now with a veneer of neutrality that makes the discrimination harder to see. This is "fairness through unawareness," and it is one of the field's best-documented failures.

FAIRNESS THROUGH UNAWARENESS FAILS

Deleting the protected attribute from the feature set does not make a model fair; it makes the model's reliance on the attribute invisible while proxies do the same work. An audit must actively test for proxy encoding—by measuring how well group membership can be predicted from the remaining features—rather than trusting that a missing column means an absent influence.

An audit interrogates proxies by asking, for the features the model uses, how strongly each predicts the protected attribute, and whether a feature's predictive value for the outcome persists once its correlation with the protected attribute is accounted for. A feature that helps mainly because it encodes group membership is a proxy doing illegitimate work; a feature that predicts the outcome for reasons unrelated to group is legitimate even if it happens to correlate with group. Distinguishing the

two is judgment-laden and cannot be fully automated, which is exactly why it belongs in a documented audit rather than a silent preprocessing step.

Missing and undersampled groups

A model performs worst where it has seen the least. If a group is undersampled in the training data, the model's errors for that group will be larger and less stable, and an aggregate accuracy metric—dominated by the majority—will not reveal it. The canonical demonstrations of this effect come from computer vision systems that performed markedly worse on the faces of women with darker skin, precisely the intersection most underrepresented in the benchmark data. The lesson generalizes far beyond vision: whenever a group is thin in the data, the audit must expect degraded and uncertain performance for it and must report performance disaggregated to that group rather than folded into an average.

Undersampling also corrupts the audit itself, not only the model. When a subgroup is small in the evaluation set, the disparity estimate for that subgroup carries wide uncertainty, and a difference that looks alarming—or reassuring—may be noise. This is why every disaggregated metric in a credible audit is reported with a confidence interval and an honest note where the sample is too small to conclude anything. A point estimate for a subgroup of a dozen people is not a finding; it is a placeholder for a finding the data cannot yet support.

Label bias: when the ground truth is not the truth

Supervised models learn to predict labels, and audits usually treat those labels as ground truth. But labels are made by people and processes, and they carry the biases of their making. A model trained to predict "arrest" is not trained to predict "crime"; arrests reflect where enforcement was concentrated as much as where offenses occurred. A model trained on past hiring decisions learns to reproduce whoever the organization hired before, biases included. A model trained on clinical costs as a proxy for clinical need will underserve groups who historically received less care, because less spending was recorded against them even when their need was equal or greater—an effect documented in a widely cited study of a health-care risk algorithm. When the label is a biased proxy for the thing the decision actually cares about, no amount of model tuning fixes the problem, and an audit that never questions the label will certify the bias as accuracy.

DATA-STAGE BIAS CHECKLIST — INTERROGATE EACH SOURCE BEFORE TRUSTING ANY METRIC

Bias source	Question to ask	Diagnostic move
Proxy encoding	Can group membership be reconstructed from the remaining features?	Predict the protected attribute from the feature set; inspect the strongest proxies
Representation gap	Which groups are thin or absent in training and evaluation data?	Tabulate sample sizes by group and intersection; flag small cells
Label validity	Is the label the real target, or a biased proxy for it?	Trace how labels were produced and by whom; identify the true quantity of interest
Measurement error	Are features or labels measured with different accuracy across groups?	Compare data quality and missingness rates by group
Historical outcome	Does the training signal encode past discriminatory decisions?	Examine whether historical decisions were themselves contested or reformed
Selection bias	Does the data over-represent one population because of how it was collected?	Compare the data-generating process to the deployment population

Documenting the data

The discipline that prevents these failures is documentation, and the field has converged on structured formats for it. A datasheet for a dataset records its provenance, its composition, the population it represents and omits, how labels were produced, and the uses for which it is and is not appropriate. An audit should demand this documentation and, where it does not exist, produce it as a first deliverable. Much of the hardest work in a bias audit is not statistical at all; it is the archaeology of understanding where a dataset came from and what its numbers actually mean. A team that cannot answer basic questions about the provenance of its training data is not ready to audit the model that data produced.

SECTION V

Metrics & Tests

Once the scope is fixed and the data understood, the audit computes. But computation is the part practitioners most often get backwards: they reach for a single fashionable metric, run it, and compare it to a round-number threshold. A credible measurement stage instead runs a small battery of complementary tests, reports each with its uncertainty, disaggregates every one across the groups in scope, and reads the results together—because each metric is blind in exactly the place another one sees.

This section is the toolkit's metric reference. It defines the core tests, states what each reveals and conceals, and gives concrete guidance on thresholds and subgroup analysis. The organizing principle is that no metric is self-interpreting: a number means something only against the confusion matrix it came from, the base rates it was computed over, and the sample size that produced it.

Start from the confusion matrix

Almost every group-fairness metric is a ratio drawn from the confusion matrix—the cross-tabulation of what the model predicted against what actually happened. An audit that computes these ratios per group, and lays the per-group confusion matrices side by side, has already done most of the interpretive work. The false positive rate and false negative rate answer "who bears the model's mistakes, and which kind." The positive predictive value answers "when the model flags someone, how often is it right—and does that differ by group." Reading these together is what distinguishes an audit from a metric dump.

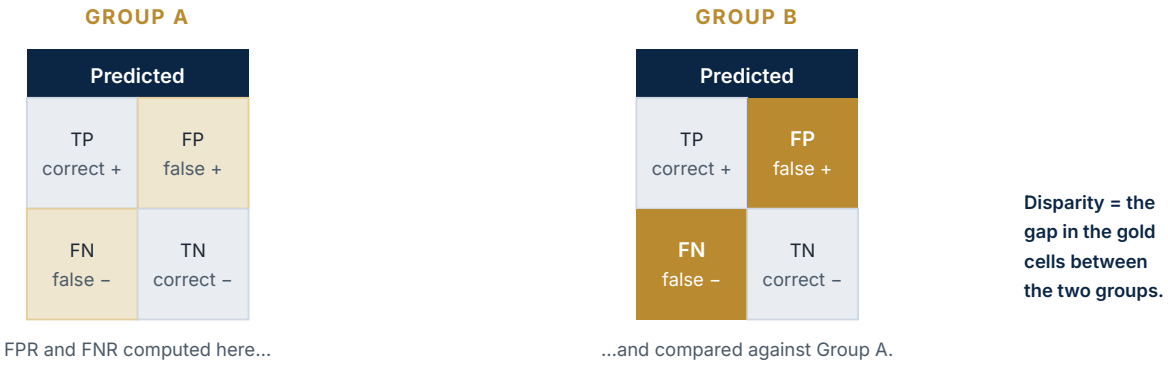

Figure 1. Group fairness reduces, at bottom, to comparing per-group confusion matrices. Equalized odds asks that the false-positive and false-negative rates (the gold cells) match across groups; other criteria read different cells of the same tables.

Error-rate parity

Error-rate parity asks whether the model's mistakes fall equally on each group. The two error rates carry different human meanings and must be examined separately. A disparity in **false negative rates** means one group is disproportionately denied the positive outcome they qualified for—missed diagnoses, overlooked qualified applicants, benefits wrongly withheld. A disparity in **false positive rates** means one group disproportionately bears the burden of the model's over-flagging—unwarranted investigations, wrongful denials, unjust scrutiny. Reporting only overall accuracy, or only one of the two error rates, conceals which harm the disparity actually inflicts. The audit reports both, per group, with intervals.

Calibration

Calibration asks whether a score means the same thing across groups. A well-calibrated score of 0.7 corresponds to a 70 percent real-world rate of the outcome for everyone who receives it, regardless of group. Calibration matters most when a human consumes the score and needs to trust it uniformly; a miscalibrated score systematically misleads the decision-maker about one group. Crucially—and this is where the impossibility result bites—a model can be perfectly calibrated and still have starkly unequal error rates. Calibration and error-rate parity are different lenses, and reporting one while implying it covers the other is a quiet way to mislead.

Disparate impact and the selection-rate view

Disparate impact looks at outcomes rather than errors: it compares the rate at which each group receives the positive decision. It is the metric most closely tied to anti-discrimination law, and it is often expressed as the ratio of the selection rate of a group to that of the most-selected group. A commonly cited rule of practice treats a ratio below four-fifths as a signal that warrants scrutiny—not a legal verdict, but a trigger for a closer look. Disparate impact has the virtue of measuring what people actually experience, and the limitation that it says nothing about whether the difference is justified by a genuine, legitimate difference in the underlying rate. It is a screening test, not a conclusion.

METRIC SELECTION GUIDE — MATCH THE TEST TO THE HARM AND THE DECISION

Metric / test	Reads	Use when the priority is...	Report alongside
False negative rate parity	Missed positives by group	No qualified person in a group is wrongly denied	False positive rate parity
False positive rate parity	Over-flagging by group	No group disproportionately bears wrongful burden	False negative rate parity
Calibration by group	Meaning of the score	A human must trust the score equally for all	Error-rate parity (they can conflict)
Disparate impact ratio	Selection rate by group	Legal exposure and experienced allocation	Justification analysis of the difference
Positive predictive value parity	Precision of a flag by group	A flag should mean the same across groups	Base rates by group
Subgroup / intersectional breakdown	Concentrated harm hidden by averages	Any audit—this is not optional	Sample sizes and confidence intervals

Subgroup analysis and the tyranny of the average

The single most valuable move in the measurement stage is disaggregation. A model can look fair on every headline metric while failing badly for a subgroup that the aggregate absorbs. Subgroup analysis deliberately breaks the population down—by group, and by intersection of groups—and recomputes every metric within each cell. It is where audits earn their keep, and it is bounded only by sample size: as the cells grow smaller, estimates grow noisier, and the audit must report where it can and cannot conclude. An audit that reports only aggregate metrics has not audited for fairness; it has audited for the comfort of the majority.

REPORT UNCERTAINTY OR REPORT NOTHING

Every disparity estimate is computed from a finite sample and carries uncertainty. A gap of a few points on a small subgroup may be noise; a smaller gap on a large one may be real and important. Confidence intervals are not decoration—they are what separate a finding from a coincidence. An audit that reports point estimates without intervals cannot distinguish signal from sampling.

Read together, these tests do something no single number can: they show the shape of a system's inequity. One system fails on false negatives for a small subgroup while looking calibrated overall; another passes disparate impact while quietly over-flagging one group. The battery reveals the

tradeoff each system has made—which, as Section II established, is unavoidable—and hands the decision-maker an honest picture instead of a reassuring scalar.

SECTION VI

The Audit Protocol

This is the operational heart of the toolkit: a repeatable, seven-stage process an audit team can run, document, and hand to an independent reviewer to run again. Every stage has an input, an output artifact, and an exit criterion.

The protocol that follows assembles the preceding sections into a single sequence. It is deliberately concrete. A good audit is not a burst of insight; it is a procedure that a competent team can execute the same way twice and that a different team can reproduce from the record. The artifacts named at each stage are the connective tissue—they are what an oversight body, a regulator, or the next auditor will actually read.

THE SEVEN-STAGE AUDIT PROTOCOL AT A GLANCE

Stage	Core activity	Exit criterion	Artifact produced
1. Scope	Fix boundary, groups, harms, fairness definition	Written scope statement approved	Scope statement
2. Data assessment	Interrogate proxies, representation, labels	Data documented; gaps flagged	Data assessment & datasheet
3. Measurement	Compute the metric battery, disaggregated	All in-scope metrics computed with intervals	Metric results tables
4. Interpretation	Read results against harms and the chosen bar	Findings stated with cause hypotheses	Findings memo
5. Remediation	Select and apply mitigations; re-measure	Chosen fixes applied and re-tested	Remediation record
6. Reporting	Document method, findings, limits, decisions	Report reviewed; residual risk stated	Audit report
7. Re-audit	Schedule and trigger repeat assessment	Cadence and triggers set and owned	Re-audit plan

Stages one through three: from scope to numbers

1 **Scope.**
Produce the written scope statement from Section III: system boundary, protected attributes and intersections, reference population, enumerated harms, the chosen fairness definition and disparity bound, and an honest note on what the data can support. Nothing downstream begins until this is approved. The exit criterion is a signed scope; the artifact is the scope statement itself.

2 **Assess the data.**
Run the Section IV checklist before touching a fairness metric: test for proxy encoding, tabulate sample sizes by group and intersection, trace how labels were produced, and document provenance in a datasheet. The exit criterion is that the data is understood and its gaps are flagged in writing; the artifact is the data assessment. A team that cannot complete this stage is not permitted to proceed on the fiction that the model is the only thing worth auditing.

3 **Measure.**
Compute the full battery from Section V—error-rate parity, calibration, disparate impact, and any metric the harms require—disaggregated across every in-scope group and intersection, each with a confidence interval. The exit criterion is a complete results table with uncertainty; the artifact is the metric results. Resist the urge to stop at the first metric that looks acceptable.

Stages four through seven: from numbers to accountable action

4 **Interpret.**
Read the numbers against the harms and the chosen bar. State which disparities exceed the tolerated bound, for whom, and how large they are relative to their uncertainty. For each material disparity, advance a hypothesis about cause—data, model, threshold, or workflow—because the cause determines the remedy. The exit criterion is a findings memo that a non-specialist decision-maker can act on.

5 **Remediate.**
Select mitigations using Section VII, matching the fix to the diagnosed cause. Apply them, then re-run the measurement stage to confirm the disparity actually fell and that the fix did not simply move the harm elsewhere. The exit criterion is applied-and-re-tested mitigation; the artifact is the remediation record, including fixes considered and rejected and why.

6 **Report.**
Document the method in enough detail to be reproduced, the findings with their uncertainty, the remediation and its effect, and—above all—the residual risk and the limits of what the audit could see. A report that omits its own blind spots is not finished. The artifact is the audit report; the exit criterion is independent review of it.

7 **Re-audit.**
An audit is a snapshot; the system moves. Set a cadence for repeat assessment and the triggers—model retraining, population shift, expanded use, a reported incident—that force an off-cycle re-audit. Assign an owner. The artifact is the re-audit plan; the exit criterion is that cadence, triggers, and ownership are all set and recorded.

THE PROTOCOL IS ONLY REAL IF IT IS REPEATABLE

The test of this protocol is not whether it produces a finding once. It is whether an independent team, handed the scope statement and the method, can rerun it on the next model version and reach comparable conclusions. Design every stage's artifact so that it survives the departure of the people who wrote it. An audit that cannot be repeated is a one-time opinion wearing the costume of a process.

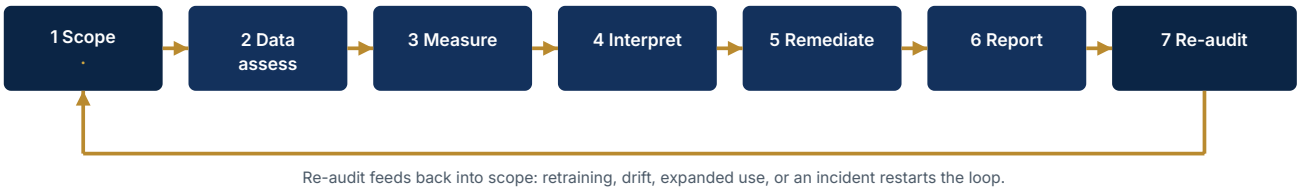

Figure 2. The audit is a loop, not a line. Stage seven returns to stage one whenever the model, the data, or the use changes—because a passing audit describes a moment, not a permanent property.

SECTION VII

Remediation

Finding a disparity is only useful if the audit can point toward a fix, and fixes come in a definite taxonomy. Technical mitigations attach to three points in the modeling pipeline—before, during, and after training—and each has characteristic strengths, costs, and failure modes. But the most important remediations are often not technical at all. A disparity produced by the workflow is not repaired by retraining the model, and a disparity rooted in a biased label is not repaired by any post-processing at all. Remediation begins with diagnosis: fix the cause you found, not the cause that is easiest to address.

The organizing insight of this section is that the three technical intervention points trade off along a common axis. The earlier you intervene—closer to the data—the more fundamental the fix and the more you must control; the later you intervene—closer to the output—the more surgical and reversible the fix and the less you need to touch. None dominates. The audit's job is to match the intervention to the diagnosed cause and to the constraints of the deployment.

MITIGATION FAMILIES — WHERE EACH INTERVENES, AND ITS TRADEOFFS

Family	Intervenes on	Best when	Cost / risk
Pre-processing	The training data—reweighting, resampling, relabeling, repair	Disparity originates in representation or historical bias in the data	Requires retraining; can distort data; needs the protected attribute at training time
In-processing	The training objective—fairness constraints or penalties	You control training and want fairness optimized jointly with accuracy	Complex; can reduce accuracy; opaque to non-specialists; ties fairness to one definition
Post-processing	The model's outputs—group-specific thresholds, score adjustment	The model is fixed or bought; a fast, auditable fix is needed	Group-specific thresholds may raise legal and ethical questions; treats symptom not cause
Workflow / human	The decision process around the model	Disparity is produced or amplified by the human step, threshold, or use	Requires organizational change; harder to measure; easy to neglect

Pre-processing: fix the data

Pre-processing mitigations act on the training data before the model ever sees it—reweighting examples so underrepresented groups carry more influence, resampling to balance thin groups, or repairing features to reduce their correlation with the protected attribute. When the diagnosed cause is a representation gap or a historical pattern baked into the data, this is the most fundamental place

to intervene, because it addresses the disparity at its source rather than compensating for it later. The costs are real: pre-processing requires retraining, it can distort the data in ways that harm accuracy if done carelessly, and it generally requires access to the protected attribute at training time. It also cannot fix a label that is measuring the wrong thing—if the target itself is a biased proxy, reweighting the examples only balances the bias, it does not remove it.

In-processing: fix the objective

In-processing mitigations change what the model optimizes, adding a fairness constraint or a penalty term so the training process pursues accuracy and a fairness criterion together. This is the most principled approach when the organization controls training, because it lets the model find the best available tradeoff rather than accepting an off-the-shelf one and patching it. Its costs mirror its power: it is technically demanding, it usually requires accepting some reduction in aggregate accuracy in exchange for reduced disparity, and—because the constraint encodes one specific fairness definition—it hard-wires the very choice Section II insisted must be made deliberately. In-processing does not escape the impossibility result; it chooses a point on the tradeoff curve and builds it into the model.

THERE IS NO FREE MITIGATION

Every technical fix has a price, usually paid in accuracy, in the need to use the protected attribute, or in the legal and ethical questions raised by treating groups differently. A mitigation that appears costless has almost certainly moved the harm somewhere the audit did not look. Always re-run the full metric battery after remediating, and check that the disparity fell without a new one rising in its place.

Post-processing: fix the output

Post-processing mitigations leave the model untouched and adjust its outputs—most commonly by setting different decision thresholds for different groups so that a chosen error rate is equalized. Their great advantage is that they work on a model you cannot retrain, including one you bought from a vendor, and they are fast, transparent, and easy to audit. Their disadvantage is twofold. They treat the symptom rather than the cause: the model still encodes whatever produced the disparity, and the adjustment merely compensates at the end. And group-specific thresholds—deciding differently for people because of their protected group—raise serious legal and ethical questions that vary by jurisdiction and domain and must be examined with counsel, not adopted casually because a library makes them easy.

Workflow and human fixes: the remedies a model cannot supply

Many audited disparities do not live in the model at all. They live in the threshold someone chose, in the order in which a screen presents candidates, in the time pressure that makes an operator defer to

the machine, or in the differential trust a decision-maker places in the score for different applicants. No technical mitigation reaches these, and an audit that recommends only model changes for a workflow-produced disparity has misdiagnosed the problem. Workflow remedies include changing how outputs are presented, adding friction or review to high-stakes decisions, monitoring override rates by group, retraining the humans, and—sometimes— concluding that the system should not be used for this decision at all. These fixes are harder to measure and easier to neglect precisely because they require organizational change rather than a code commit, which is exactly why the audit must name them explicitly.

> *The cheapest durable fix is usually not a new model. It is a better threshold, a cleaner label, or a redesigned human step.*
>
> — On matching the remedy to the cause

A **Diagnose before you prescribe.**
Trace each material disparity to its cause—data, objective, threshold, or workflow—before selecting a mitigation. The cause dictates the family of fix; skipping diagnosis wastes effort on the wrong layer.

B **Prefer the fix closest to the cause.**
A representation gap wants pre-processing; a workflow effect wants a workflow change. Compensating for a data problem at the output leaves the cause in place to resurface at the next retraining.

C **Re-measure and watch for displaced harm.**
After any mitigation, rerun the full battery. Confirm the target disparity fell and that no new disparity rose for another group or on another metric. Fairness fixes routinely move harm rather than remove it.

D **Know when the answer is not to deploy.**
Some disparities cannot be remediated to an acceptable level for the stakes involved. A recommendation not to use the system, or to restrict its use, is a legitimate and sometimes necessary audit outcome.

SECTION VIII

Governance

An audit that is not documented, not repeated, and not independent is not assurance—it is a favorable opinion the organization paid for. Governance is what turns a one-time exercise into a standing commitment that outlives the people who ran it.

The preceding sections describe how to run an audit. This one describes how to make audits a durable part of how an organization operates: how to document them so they can be checked, how to set the thresholds that convert measurements into decisions, how often to repeat them, and how to secure the independence without which none of it is credible. These are not add-ons to the technical work; they are what give the technical work its weight.

Documentation: the audit is its record

An audit exists, for all practical purposes, only insofar as it is written down. The field has converged on structured documentation formats for exactly this reason. A model card records a model's intended use and its performance disaggregated across groups and conditions. A datasheet records a dataset's provenance and composition. An audit report ties these together with the scope statement, the metric results and their uncertainty, the remediation and its effect, and an explicit statement of residual risk and blind spots. The test of good audit documentation is reproducibility: could an independent party, given the record, rerun the audit and reach the same conclusions? If not, the documentation is incomplete, however impressive its charts.

4

standing governance instruments: documentation, thresholds, cadence, and independent re-auditing

3

documentation artifacts an audit ties together: model card, datasheet, and audit report

2

re-audit triggers beyond the calendar: material model change and reported harm

Thresholds: turning measurement into decision

A metric is not a decision until a threshold is attached to it, and the threshold is a governance choice, not a technical one. How much disparity is too much? A four-fifths selection-rate ratio is a familiar screening trigger, but the acceptable bound for any given metric depends on the stakes, the reversibility of the harm, and the law—and the audit cannot set it unilaterally in the moment of measurement without inviting the suspicion that the bar was drawn to fit the result. The disciplined practice is to set thresholds in advance, in the scope statement, so that the audit measures against a bar it committed to before it knew the answer. Thresholds set after the fact are not standards; they are rationalizations.

GOVERNANCE CHECKLIST — THE STANDING COMMITMENTS THAT MAKE AUDITS CREDIBLE

Commitment	What good looks like	Warning sign
Documentation	Scope, method, results with uncertainty, remediation, and residual risk are recorded and reproducible	Findings live in a slide deck no one can rerun
Thresholds	Disparity bounds set in advance, tied to stakes and law, in the scope statement	The acceptable bound is decided after seeing the result
Cadence	Re-audits scheduled at a fixed interval, with defined off-cycle triggers	The system is audited once, at launch, then trusted indefinitely
Independence	The auditor is organizationally separate from the team that built or champions the system	The building team grades its own homework
Ownership	A named person is accountable for the audit program and empowered to act on findings	Responsibility is diffuse; findings have no owner
Transparency	Findings and methods are disclosed to oversight bodies and, where appropriate, the public	Results are held closely and never externally checked

Cadence: an audit is a snapshot, not a certificate

A model retrains, a population shifts, an upstream feed changes format, a tool's use quietly expands—and the fairness established at one audit decays. Governance requires a cadence: a scheduled re-audit at a fixed interval, plus explicit triggers that force an off-cycle assessment whenever the system materially changes or a harm is reported. The interval should scale with the stakes; the highest-consequence systems warrant the most frequent re-examination. What is not acceptable is the common pattern of a single audit at launch, filed and forgotten, certifying in perpetuity a system that has since changed underneath its conclusions.

INDEPENDENCE IS THE WHOLE GAME

An audit graded by the team that built the system is not an audit; it is a defense brief. Credible assurance requires that the auditor be organizationally separate from—and not beholden to—the people who want the system deployed. Independence can come from an internal unit with a distinct reporting line or from an external party, but it cannot come from the deploying team no matter how well-intentioned. Self-graded homework is not evidence.

Independent re-auditing and external accountability

The strongest form of assurance is an audit that an outside party can check. Internal audits are necessary and valuable, but they operate under organizational pressures that even scrupulous teams feel. Independent re-auditing—by a separate internal unit, an external firm, or a regulator—tests

whether the findings survive scrutiny by someone with no stake in the answer. This is where the documentation discipline pays off: an audit built to be reproducible can be handed to an independent party and rerun; an audit that lives in one team's private tooling cannot. The direction of the field, in both emerging regulation and mature practice, is toward exactly this kind of external accountability, and an organization that builds its audits to be independently reproducible is building for where the standards are going.

CLOSING NOTE

A bias audit is not a verdict of fairness and was never going to be. It is a disciplined way to see where a system distributes harm unequally, to say so honestly under stated assumptions, and to be held to what it found. Run that way—scoped deliberately, measured with uncertainty, remediated at the cause, documented for reproduction, and repeated independently—an audit does the one thing a fairness slogan never can: it makes the treatment of the people a system touches an accountable, checkable fact rather than a hopeful claim.

APPENDIX

Notes & Further Reading

This toolkit synthesizes established results and practice in algorithmic fairness, measurement, and AI documentation. The sources below are entry points into the literature that informs it; they are indicative rather than exhaustive.

1. Barocas, S., Hardt, M., & Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities.* A comprehensive treatment of fairness definitions, tradeoffs, and their sociotechnical context.
2. Chouldechova, A. *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.* On the impossibility of simultaneously satisfying calibration and equalized error rates when base rates differ.
3. Kleinberg, J., Mullainathan, S., & Raghavan, M. *Inherent Trade-Offs in the Fair Determination of Risk Scores.* A formal statement of the incompatibility among leading group-fairness criteria.
4. Hardt, M., Price, E., & Srebro, N. *Equality of Opportunity in Supervised Learning.* On equalized odds and post-processing methods for enforcing it.
5. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. *Fairness Through Awareness.* On individual fairness and the limits of fairness through unawareness.
6. Buolamwini, J., & Gebru, T. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.* On disaggregated, intersectional evaluation and the harm of undersampled groups.
7. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.* On label bias arising from a proxy target.
8. Mitchell, M., et al. *Model Cards for Model Reporting.* On structured documentation of model performance disaggregated across groups and conditions.
9. Gebru, T., et al. *Datasheets for Datasets.* On documenting dataset provenance, composition, and appropriate use.
10. Raji, I. D., et al. *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing.* On repeatable, documented audit processes across the development lifecycle.
11. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. *Fairness and Abstraction in Sociotechnical Systems.* On why fairness cannot be resolved at the level of the model alone.
12. National Institute of Standards and Technology. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (NIST SP 1270)* and the *AI Risk Management Framework.* On categories of AI bias and structured approaches to managing them.