A 42 CONSULTING.AI WHITEPAPER

The Fairness Threshold Problem

Why Fairness Is a Threshold Someone Must Choose, Not a Property a System Has

Dr M Maruf Hossain, PhD, GAICD
Chief AI Strategist · 42 Consulting.AI
enquire@42consulting.ai · www.42consulting.ai

42 Consulting.AI
Scaling Intelligence

CONTENTS

Table of Contents

FOREWORD

From the Author.

Ask whether an AI system is fair, and you will usually get a number back, delivered with the confidence of a measurement: *Is it fair? Yes — the disparity is within tolerance.* But the number conceals a decision that someone made, and few boards ever see: which fairness, measured how, and held to what line. Change the metric, and the same system passes or fails. Fairness was not found in the data. It was chosen.

This is uncomfortable because we would prefer fairness to be a property a system either has or lacks — something an engineer certifies and a board ticks. It is not. There are several mathematical definitions of fairness; they capture different intuitions; and in the ordinary case where groups differ in their base rates, they are provably incompatible — satisfying one means breaching another. No configuration is fair on every definition at once. Someone has to decide which fairness the organisation will stand behind, and at what threshold.

This paper is about that decision, and about why it cannot be left implicit. It is the difference between a system whose fairness was chosen, owned, and defended, and one whose fairness is whatever the default metric happened to report. When a regulator or a declined applicant asks whether a decision was fair, the organisation that chose its threshold can answer. The organisation that never chose is left defending a number it does not understand.

It is written for the board and the executive who have been shown a fairness dashboard and told that the system "passed". Passed which test, against whose threshold, chosen by whom? This paper offers no formula that dissolves the problem — none exists. It shows where the choice lives, who must own it, and why the threshold, once chosen, has to be a barrier the system cannot cross rather than a number a delivery team can quietly relax.

M Maruf Hossain

SECTION ONE

Executive Summary.

Fairness in AI is not a single property with a single measurement. It is a family of incompatible definitions, and choosing among them is a governance act, not a technical one. Most organisations never make the choice explicitly — they inherit whatever their tooling defaults to. This whitepaper establishes three propositions.

First, there is no single fairness — its definitions conflict. Demographic parity, equalised odds, and counterfactual fairness each capture a real and different intuition about what fair treatment means. When groups differ in base rates — as they almost always do — these definitions are mathematically incompatible: a system that satisfies one will violate the other. "Make it fair" is therefore an under-specified instruction. It cannot be carried out until someone names which fairness is meant.

Second, the threshold is a choice someone must own, not a fact to be discovered. Because no metric is free of trade-offs, the level at which a disparity becomes unacceptable is a decision, not a reading off the data. It should be set deliberately — tighter where the system's authority and the stakes are higher — and it should carry a name: the person accountable for choosing it. A fairness threshold with no owner is not a standard; it is an accident waiting to be defended after the fact.

Third, bias past the threshold is a barrier, not a risk to accept. Once the threshold is chosen, a system that exceeds it is not allowed to proceed. Predictive performance does not buy the right to tolerate a discriminatory outcome for a protected group, and "accept the bias risk" is not a remediation. Where bias cannot be brought within the chosen threshold, the responsible outcome is that the system does not deploy — or is withdrawn if it already has.

This paper sets out why fairness reduces to a choice, names the three metrics and the conflict between them, shows how the threshold is set and by whom, describes the protocol for cases where metrics collide, and shows the architecture that makes the threshold a gate rather than a suggestion. It ends with five questions a board should be able to answer about the fairness it chose — not next quarter, but today.

SECTION TWO

More Than *One Fairness.*

Ask what it means for an automated decision to be fair, and more than one reasonable answer appears. One says the system should approve people from each group at the same rate — equal outcomes across groups. Another says that among those who actually qualify, the system should approve at the same rate regardless of group, and make its mistakes at the same rate too — equal accuracy across groups. A third says the decision for an individual should not change if you alter only their protected characteristic and nothing else — no dependence on group at all. Each is a genuine account of fairness. They are not the same account.

These correspond to the three metrics the framework requires: demographic parity, equalised odds, and counterfactual fairness. Each is precisely defined and separately measurable. The difficulty is not that any one of them is wrong. It is that they pull in different directions, and the tension is not a shortfall of engineering effort that a better model could close. It is a property of mathematics.

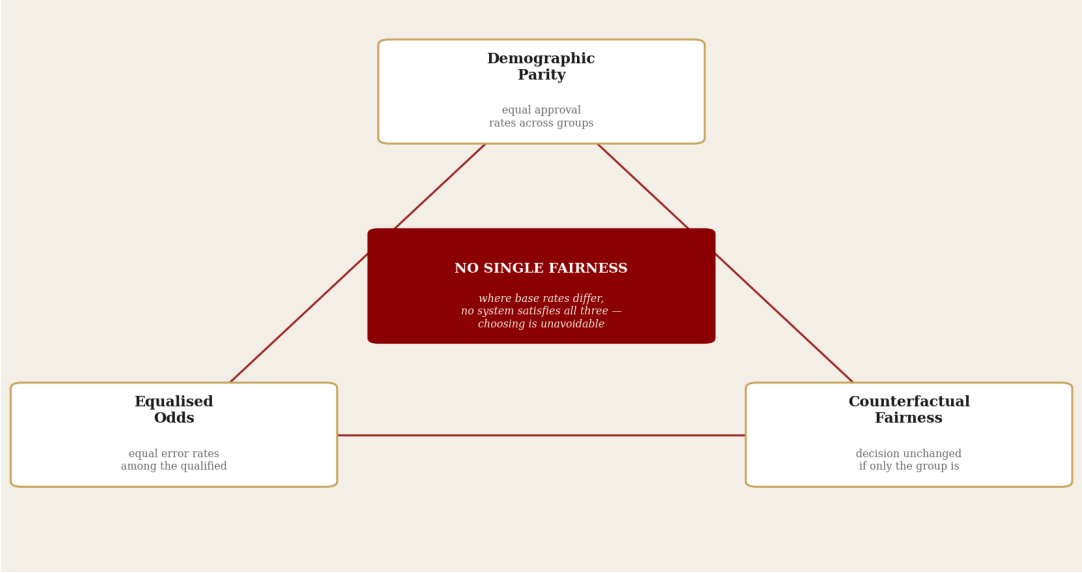

Figure 1 — Three definitions of fairness, one system. Where base rates differ across groups, the metrics cannot all be satisfied at once; improving one moves another.

The incompatibility is exact. Where two groups have different underlying base rates — different proportions who actually qualify, for reasons the model did not create — it is provably impossible to satisfy both demographic parity and equalised odds simultaneously. Forcing equal approval rates across groups distorts the error rates; equalising the error rates reopens a gap in approval rates. This is not a defect to be patched but a theorem, and no amount of retraining escapes it (Hossain, 2026).

So a system cannot be "fair" in the unqualified sense the word invites. It can be fair on a named definition, to a stated tolerance. When someone reports that a model "passed its fairness test", the only meaningful questions are which definition and what threshold. A pass on demographic parity says nothing about equalised odds — a system can look impeccable on one while producing exactly the disparity the other was meant to catch.

The three definitions are set out below. Each is precisely defined and separately measurable; none is wrong. What the table makes plain is that they are not interchangeable, and that satisfying one forecloses another. The final column is the decision this paper argues cannot be left to a default.

Fairness definition	What it requires	What it cannot also guarantee	Where it is the right primary metric
Demographic parity	Approval rates equal across groups, regardless of who qualifies	Equalised odds, wherever base rates differ between groups	Access to opportunity, where equal representation is the stated policy objective
Equalised odds	Equal error rates among those who actually qualify, regardless of group	Demographic parity, wherever base rates differ between groups	Decisions where a wrong answer harms the individual: credit, fraud, eligibility
Counterfactual fairness	The same outcome had the individual's protected characteristic been different	Either of the above, in the general case	Decisions turning on individual merit, where proxy discrimination is the risk

Table 1 — Three definitions of fairness, and what each forecloses.

Read across any row, and the trade is visible. Read down the third column, and the reason a single unqualified claim of fairness cannot be made becomes clear. The governing definition is a choice about which harm the organisation will not accept, made before the metric is computed rather than after it is reported.

This is why fairness cannot be delegated to a metric and forgotten. The choice of which definition governs a given system is the substantive decision, and it belongs to the people accountable for the system's effect on individuals — not to whichever metric a library happened to compute by default. The rest of this paper is about making that choice deliberate, owning it, and holding the line it sets.

SECTION THREE

The Threshold *is a Choice.*

If fairness must be measured on a named definition to a stated tolerance, then the tolerance — the threshold — is where the real governance happens. A threshold is the maximum disparity the organisation is prepared to accept on a given metric, for a given protected characteristic, before the outcome counts as unfair. It is not read from the data. It is set. And the level at which it is set is a value judgement dressed, too often, as a technical parameter.

The framework refuses to leave that judgement implicit. It fixes thresholds in advance, by system tier, and makes them tighter where the consequences are graver. A system deciding credit, employment, or access to essential services is held to the strictest tolerances — a few percentage points of disparity, per characteristic, not averaged away across a portfolio. A lower-stakes internal system is held to a looser line. The principle is proportionality: the greater a system's authority over people's lives, the higher the fairness performance it must demonstrate before it is allowed to act.

Two features of this design matter. First, the threshold is per-characteristic and absolute: a system that meets its tolerance on every characteristic but one has still failed, because a disparity that harms a protected group is not offset by fairness elsewhere. Second, the threshold is a governance artefact with an owner. Setting it — and any decision to hold a system to a looser line than its tier would suggest — is an accountable act recorded against a name, not a default a delivery team can quietly widen when a model will not otherwise pass.

None of this pretends the numbers are objective truths. The framework is explicit that the thresholds are deliberate design choices, calibrated to proportionality and open to being tightened where regulation or sector demands. What it insists on is that the choice be made openly, by someone with the authority to make it, and written down — so that when the organisation is asked why this level and not another, the answer is a decision it can defend, not a parameter no one remembers selecting.

SECTION FOUR

When the Metrics *Collide.*

Fixing thresholds does not dissolve the conflict between metrics; it makes the conflict explicit and forces a decision when a system meets one and breaches another. The framework provides a defined protocol for exactly that case (Hossain 2026) — not to decide whether the system is deployed, but to ensure the decision is made deliberately, documented, and disclosed. It turns an implicit trade-off into an accountable one. Its steps run as follows.

Name the primary metric before the conflict, not after. For each system, the governing fairness metric is designated in advance, based on its decisions and the errors that matter most in that domain. A metric chosen after the results are in is a metric chosen to pass. Fixing it beforehand prevents the trade-off from being resolved in whichever direction is convenient once the numbers are known.

Document the conflict in full. Where the system satisfies its primary metric but breaches a secondary one, the tension is recorded explicitly: which metric was met and by how much, which was breached and by how much, and why the primary was the right one to govern this decision. The trade-off is written down as a trade-off, not buried inside an aggregate score that reports only the comfortable number.

Escalate the decision to the authority that owns the stakes. The protocol does not let the assessor quietly resolve the conflict. The decision to proceed despite a secondary breach rises to the governance level appropriate to the system's authority — and for the highest-authority systems, the threshold itself must be set by the board before assessment even begins. Who decides is determined by how much is at stake, not by who happens to be running the test.

The point of the protocol is not to make the conflict disappear — it cannot be made to disappear. It is to ensure that when fairness definitions collide, a named person chooses which one governs, records why, and answers for it. A system that passes because someone silently picked the flattering metric is not fair; it is unexamined. The protocol is what converts an unavoidable trade-off into a defensible decision.

SECTION FIVE

The Gate That *Holds.*

A chosen threshold is only as real as the organisation's inability to ship past it. The framework therefore makes the fairness assessment a mandatory gate: no system affecting individuals reaches production without a current assessment on record, and the assessment resolves into one of four outcomes, each with a fixed consequence.

Pass — within threshold on every characteristic. The system's disparities fall within the chosen tolerance for each protected characteristic and applicable metric. It proceeds, and its fairness status is recorded so the currency of that pass can be tracked over time rather than assumed.

Conditional pass — a minor, remediated deviation. A small breach, with an approved remediation plan, an interim monitoring arrangement, and a named deadline. It proceeds under conditions that are obligations, not aspirations — and a reassessment date the system cannot outrun.

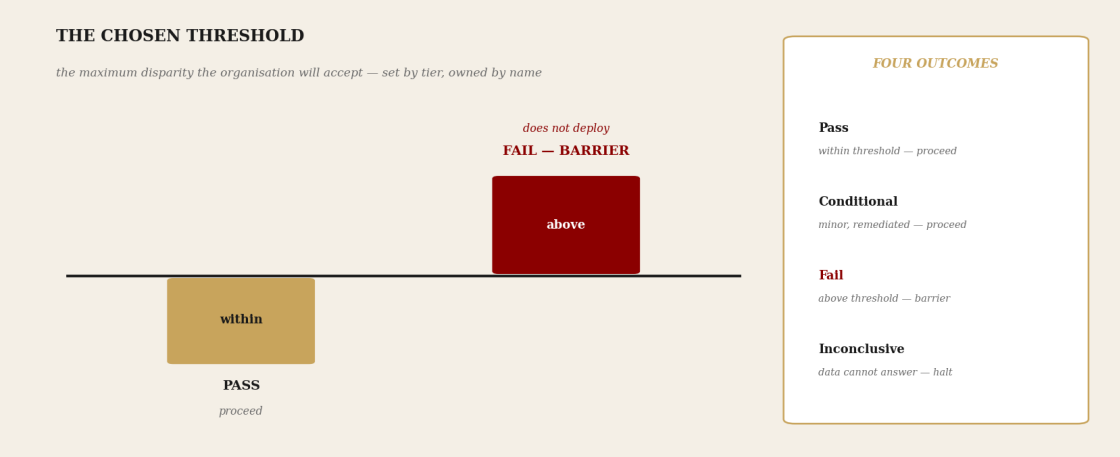

Figure 2 — Fairness as a deployment gate. Only a system within its chosen threshold passes; bias above the line is a barrier, not a residual risk.

Fail — bias above the threshold. A material breach, or a minor one with no approved plan, is a deployment barrier. The system does not proceed; the bias is recorded as a nonconformity, and remediation — retraining, redesign, scope restriction, or withdrawal — is mandatory. Accepting the bias is not among the options because accepting bias above the threshold is not remediation.

Inconclusive — the data cannot answer the question. Where there is insufficient data to assess a required characteristic, the system does not proceed on the basis of a gap. A data-collection or proxy-assessment plan must be approved, and the assessment completed, before the gate opens. Absence of evidence is not evidence of fairness.

SECTION SIX

Bias Is Not *a Risk to Accept.*

Everything so far depends on one structural commitment: that bias above the chosen threshold is a barrier, not a risk to be weighed against other objectives. It is the same distinction that separates a boundary from a negotiation, and it is where fairness regimes most often quietly fail.

The failure is familiar. Framed as a risk, a fairness breach becomes something to be "managed" — assigned an owner, given a rating, and, under delivery pressure, accepted. A model that will not pass is shipped anyway because its accuracy is valuable, its launch is due, and the disparity is logged as a residual risk that someone has signed off on. Each step looks reasonable; the sum is a system that discriminates against a protected group, with a paper trail that calls it acceptable. The threshold was real until the moment it became inconvenient.

The framework forecloses that move by design. Predictive performance is explicitly refused as a justification for tolerating discriminatory outcomes: a system's aggregate accuracy does not buy the right to be unfair to the group it disadvantages. Bias above the threshold is not a cost to be offset by benefit; it is a line the system may not cross, and the governing policy states it in exactly those terms — accepting bias above the threshold is not remediation, and a system that cannot be brought within the line may not be deployed in domains that affect individuals.

So the fairness threshold is not held in a delivery team's risk register. It is a deployment gate that the architecture enforces, and a breach is a nonconformity with mandatory remediation, not a box on a form. This is not indifference to commercial reality. It is the recognition that a fairness standard which bends under pressure is not a standard — and that the only threshold worth choosing is one the organisation has put beyond the reach of the people who would find it convenient to move.

SECTION SEVEN

Owned, *Not Assumed.*

A fairness threshold that no one chose, and a fairness assessment no one can produce, are the two ways the whole edifice collapses back into assumption. The framework closes both by wiring fairness into the same machinery that governs the rest of the AI estate, so that the choice is recorded, the evidence is retained, and neither can be quietly skipped.

The line runs from constitution to evidence. Fairness and non-discrimination are adopted at the board level as a governing principle; the standard defines the metrics and thresholds; the policy makes the assessment a mandatory gate; and the results are recorded in the registers that track each system and surfaced to the board through a small set of indicators. Choosing the threshold, passing the gate, and remediating a breach are all recorded acts with owners — not assurances offered after the fact.

Crucially, fairness attaches to the existing enforcement spine rather than sitting beside it as good intentions. The deployment gate that clears a system is extended to require a current, passed fairness assessment. The register that records a system's authority carries its bias status and the date its assessment lapses. The people who supervise systems in operation are briefed to recognise unfair outcomes, not only breaches of operating limits. Fairness becomes a precondition the system had to satisfy, checked and monitored — not a claim in a policy.

The effect is that fairness becomes an artefact rather than an assurance. When a regulator asks which fairness the organisation chose and why, the answer is the designated metric, the threshold, and the name of whoever set it. When asked whether a system is fair today, the answer is the current assessment and its outcome, not last year's recollection. A fairness that can be shown is a fairness that was chosen and owned; one that can only be asserted was always just the default nobody examined.

None of this makes fairness simple — the conflict between definitions is permanent, and the thresholds are judgements reasonable people will contest. What the architecture provides is not the right answer but an owned one: a choice made deliberately, by someone accountable, recorded where it can be questioned, and enforced as a barrier the system cannot cross. That is the difference between an organisation that can defend its fairness and one that merely hopes the number holds.

SECTION EIGHT

Five Questions *for Your Board.*

The measure of a fairness regime is not the sophistication of its metrics. It is whether the board can show that a definition was chosen, a threshold was set by someone who owns it, and a system that breached it did not ship. Each of the following is a question a regulator, a court, or a declined individual could put to the board — and each should be answerable today, with an artefact rather than an assurance.

Which fairness did we choose — and who chose it? For each consequential system, can you name the primary fairness metric it is governed by, the threshold it must meet, and the person accountable for that choice — or did the organisation inherit whatever its tooling reported?

Is the threshold proportionate to the stakes? Are higher-authority systems — those deciding credit, employment, or essential services — held to tighter fairness tolerances than low-stakes ones, per protected characteristic and not averaged across the portfolio?

What happens when the metrics conflict? When a system passes its primary fairness metric but breaches a secondary one, is there a defined protocol that documents the trade-off and escalates the decision to an accountable authority — or is the conflict resolved silently in favour of the flattering number?

Is bias above the line a barrier, or a risk someone can sign off? When a system exceeds its fairness threshold, does it fail to deploy and enter mandatory remediation — or can a manager, under a deadline, accept the bias as a residual risk because the model performs well? Can you show a case where the answer was no?

Can we show the fairness, not just assert it? If asked whether a system is fair today, could you produce the current assessment, the chosen metric, and the threshold — or only the recollection that it passed something once? A framework that cannot answer these questions has not chosen its fairness. It has assumed it.

ABOUT

42 Consulting.AI.

42 Consulting.AI builds constitutional AI governance for regulated enterprises. Its core work, the MANDATE Suite, is a comprehensive governance architecture comprising 64 normative instruments across five frameworks — governing the delegation of machine decision authority, the ethical and rights boundaries on its use, the management system that maintains it, and the regulatory extensions for EU and group structures. The Suite is designed for organisations in financial services, insurance, superannuation, healthcare, government, and other regulated sectors operating autonomous AI systems at scale.

The architecture is supported by an academic preprint series published on Zenodo, which establishes the theoretical basis for each governance mechanism, and by the forthcoming book *Governance-First AI*. The Suite's distinguishing conviction is that the AI governance that regulated enterprises require cannot be delivered by a compliance framework alone — it requires a constitutional architecture, adopted by the board, enforced by the infrastructure, and continuously tested against current conditions.

Dr M Maruf Hossain, PhD, GAICD · Chief AI Strategist
enquire@42consulting.ai · www.42consulting.ai

The preprint series is available at: zenodo.org/communities/mandate/records

REFERENCES

References *and Sources.*

The regulatory instruments referenced in this paper were verified at the primary source. The governance mechanisms are set out in full in the MANDATE preprint series, published through Zenodo, and in the forthcoming book Governance-First AI. The full MANDATE preprint series is available at: zenodo.org/communities/mandate/records

REGULATORY AND INSTITUTIONAL SOURCES

European Union (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), Articles 9 and 10 — risk-management and data-governance obligations for high-risk AI, including examination and mitigation of bias.

ASIC (2015). Regulatory Guide 271: Internal Dispute Resolution — RG 271 obligations bearing on equitable and accountable automated decisions affecting consumers.

Commonwealth of Australia. Racial Discrimination Act 1975; Sex Discrimination Act 1984; Disability Discrimination Act 1992; Age Discrimination Act 2004 — the protected characteristics against which AI outcomes must not systematically discriminate.

Office of the Australian Information Commissioner. Privacy Act 1988 (Cth), Australian Privacy Principles — collection, use, and profiling limits bearing on the data used to train and assess AI systems.

United Nations (2011). Guiding Principles on Business and Human Rights — human rights due diligence, applicable to discriminatory outcomes from automated decisions.

BOOK

Hossain, M Maruf and Ponnusamy, Ahilan (2026). Governance-First AI. Apress (forthcoming) — the primary public articulation of the responsible AI governance doctrine.

MANDATE PREPRINT SERIES

Hossain, M Maruf (2026). The Fairness Threshold Problem: A Governance Framework for Resolving Metric Conflicts in Enterprise AI Fairness Assessment. Preprint. Zenodo. https://doi.org/10.5281/zenodo.20728389.

Consulting.AI

Scaling Intelligence

enquire@42consulting.ai · www.42consulting.ai