A 42 CONSULTING.AI WHITEPAPER

The Responsibility Gap

Why a System Can Pass Every Control and Still Harm the People It Affects

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.

Most boards ask one question of their AI governance, and it is a reasonable one to ask first. *Is the system under control?* Does it stay within the authority we granted it, escalate when it should, and stop when it is told to? A framework that answers that question well is a real achievement, and organisations are right to want it. But it is not the only question, nor is it the one that determines whether the organisation has behaved responsibly.

There is a second question, and it is about the person on the other side of the decision. Not whether the system stayed within its limits, but whether — within those limits — it treated that person fairly, could explain itself to them in terms they could understand, and was built on data that someone could actually vouch for. A system can satisfy every control on the first question and fail the second one completely. The two are simply not the same test.

This paper is about that second question, and about the gap that opens when a governance framework answers only the first. It is the distance between a system that is well governed and one that is used responsibly — and it is the space a regulator, a court, or a harmed customer will stand in when they ask the organisation to account for an outcome. A clean control record is no answer to the question they are actually asking.

It is written for the board that has been assured its AI is "governed" and wants to know whether that assurance reaches the people the AI affects. It does not resolve the question with more principles. It shows where the question lives, and what an architecture that takes it seriously is obliged to do.

M Maruf Hossain

SECTION ONE

Executive Summary.

Controlling an AI system and using it responsibly are two different obligations. Most AI governance answers the first and quietly assumes it has answered the second. It has not. This whitepaper establishes three propositions.

First, governing authority is not the same as governing impact. Whether a system stays within the limits it was given is a question of the organisation's control over its own machinery. Whether that system is fair, explicable, and built on sound data is a question about its effect on the people it touches. These are different questions with different evidence and different failure modes. A framework can be complete on the first and silent on the second, and most are—they were designed to keep the machine in its lane, not to answer for what happens to the person in the next one.

Second, responsibility is owed to the people affected, not to the framework. Accountability for an AI decision does not end with the committee that sanctioned the system; it extends to the individual on the receiving end of the decision and is measured by the outcome they experience rather than by the process that produced it. A declined applicant, a mischaracterised customer, a person who cannot get an explanation — none of them is answered by the existence of a policy. Responsibility is discharged in outcomes, not in governance artefacts about outcomes.

Third, the boundaries of responsible use are deployment barriers, not risks to accept. Unjustified harm, unfair outcomes across protected characteristics, decisions that cannot be explained to those they affect, and data of unknown origin are not residual risks to be logged and tolerated at management's discretion. Where they are present and cannot be resolved, the responsible outcome is that the system does not proceed. Treating these as barriers rather than as risks to be managed is the single structural difference between a framework that protects people and one that documents them.

This paper sets out the second question in full, explains why authority-centred frameworks miss it, describes the four boundaries that answer it, and shows the architecture that enforces them rather than merely asserting them. It ends with five questions a board should be able to answer about the people its AI affects — not next quarter, but today.

SECTION TWO

Two Questions, *Not One.*

Every AI system in a regulated enterprise raises two governance questions that are easily mistaken for one. The first is whether the system operates within its authority — whether the decisions it takes fall inside the limits the organisation delegated to it, and whether it escalates or halts when those limits are reached. The second is whether, operating entirely within that authority, the system is used responsibly — whether it is fair to the people it affects, explicable to those who must understand it, and built on data whose provenance is known.

The two questions are independent, and that independence is the whole point. A credit model can stay perfectly within its delegated authority — never exceeding its mandate, never acting outside its bounds, escalating exactly when its rules require — and still decline applicants along lines it was never meant to draw. The authority question returns a clean answer. The responsibility question returns a very different one. Nothing about staying within the limits guarantees fairness, explicability, or data quality of what happens within them.

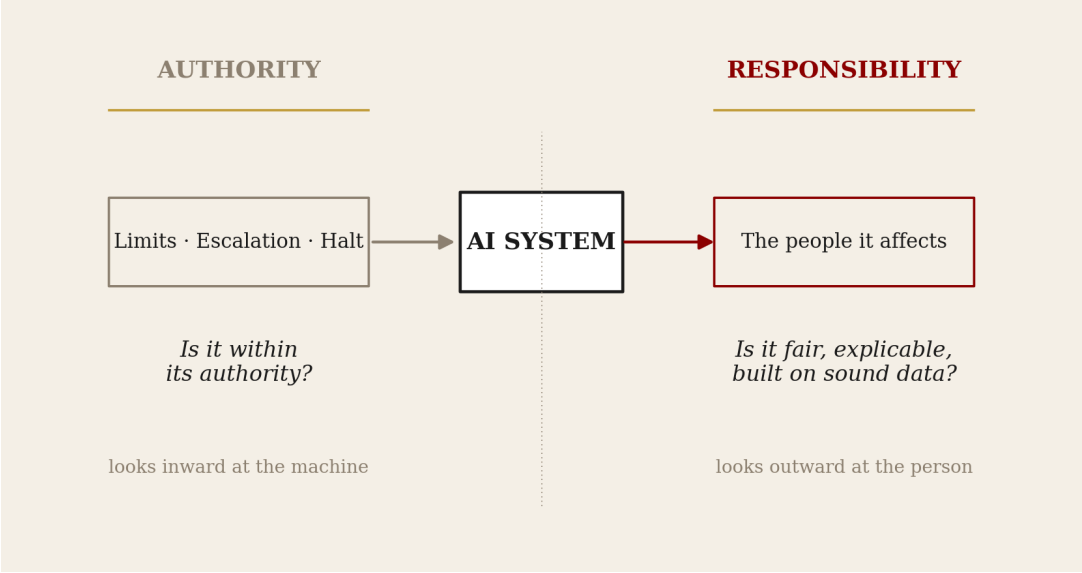

Figure 1 — Two questions, one system. Authority governance looks inward at the machine's limits; responsibility governance looks outward at the people the machine affects.

This is why an organisation can hold an unbroken record of control and still be unable to answer for an outcome. The governance that watches the machine is looking inward — at delegation, limits, escalation, and halt conditions. The governance that must answer for the outcome looks outward — at the person the decision landed on, and at whether that person was treated in a way the organisation can defend. An enterprise that has built only the inward-facing half has built half a governance system and called it whole.

The Responsible AI Framework exists to ask the outward-facing question. It is a peer to the framework that governs authority, not a subordinate part of it: a system can be flawless under one and in breach of the other, and each failure is real. The rest of this paper is about what the outward question demands, and why an organisation that answers only the inward one is exposed precisely where it believes it is covered.

This is why accountability cannot be answered by pointing to a governance committee or a policy document. A committee is not an accountable person. A policy is not a record of authorisation. When a regulator traces the chain, it arrives at an individual and asks what that individual authorised and how they oversaw it. The framework either equips that person with an answer or it does not.

SECTION THREE

The People the *System Affects.*

The reason the second question matters is that accountability for an AI system does not stay inside the organisation. It reaches the people the system acts upon. When an automated decision denies someone credit, mischaracterises them to their disadvantage, exposes them to a safety risk, or subjects them to a demeaning interaction, the organisation is responsible for that consequence — regardless of who built the model, who supplied the data, or how faithfully the system stayed within its limits.

Harm, in this context, is not a single thing. It is financial — loss, debt, the denial of a service someone needed. It is reputational — an incorrect adverse characterisation that follows a person. It is physical, where decisions touch safety-critical domains. It is psychological, in distressing or demeaning automated treatment. And it is social, where decisions quietly reinforce an existing disadvantage. A framework built to keep a system within its authority is not looking for any of these, because none of them is a breach of authority. They are breaches of responsibility.

This is the shift the Responsible AI Framework makes: from governing the organisation's control of its machinery to governing the machinery's effect on people. The obligation cannot be transferred. It does not move to the vendor who built the model, the operator who ran it, or the individual who was harmed. The organisation that deploys a system owns the consequences of that deployment. "We did not build it" and "the system stayed within its limits" are, to the person on the receiving end and to the regulator behind them, not answers at all.

Once responsibility is understood as something owed outward, the governance task changes shape. It is no longer enough to prove the organisation controlled its systems. The organisation must be able to show that, within that control, it treated people fairly, explained itself to them, and knew the provenance of what it acted on. Those are four distinct obligations, and the next sections take them in turn — first by showing where authority governance leaves them unaddressed, then by naming the boundaries that address them.

SECTION FOUR

Where Authority *Governance Ends.*

The clearest way to see the gap is to look at systems that are entirely within their authority and still indefensible in their impact. Each of the following is a governance failure. None of them is a breach of authority.

Within authority, but unfair. A model operates exactly as delegated and produces outcomes that differ across protected characteristics in ways the organisation cannot justify. No limit was exceeded. But there is no answer to the question of whether the disparity is acceptable because no one was ever required to choose a fairness threshold, test it against, and own the choice. The authority framework never asks, so the disparity ships.

Within authority, but unexplainable. A system declines a person, and the organisation can reconstruct — for an auditor, in technical terms — how the model reached its output. It cannot give the person a reason they could understand and act on. The system is explainable to the people who built it and opaque to the people it affects. Under an authority framework that treats a technical audit trail as sufficient, this passes. Under any obligation owed to the affected individual, it does not.

Within authority, but built on unknown data. A model runs inside its limits on data whose origin, quality, and consent basis no one has established. The authority governance has nothing to say about this, because provenance is not an authority question. Yet the model's biases, its blind spots, and its legal defensibility are all downstream of that unexamined data. Unknown data does not stay unknown; it surfaces later, at scale, in the outcomes people experience.

In each case, the inward-facing governance returns a clean report, and the outward-facing question is never reached. This is not a matter of poorly implemented frameworks. A perfectly implemented authority framework produces exactly this result, because these outcomes fall outside what it was built to see. Closing the gap requires a second framework that is built to see them — and that framework begins by naming four boundaries.

SECTION FIVE

Four Principles, *Four Barriers.*

The Responsible AI Framework answers the outward question through four governing principles. They are neither aspirations nor weighted against commercial considerations. Each is a boundary on use: where it is breached and cannot be remedied, the system does not deploy or ceases to operate.

Human dignity and non-harm. An AI system must not be used in ways that demean, discriminate against, or cause unjustified harm to individuals or groups. This boundary is non-delegable and cannot be waived by operational necessity. A material risk of unjustified harm is a barrier to deployment, not a risk to be entered in a register and accepted.

Fairness and non-discrimination. Outcomes must be fair across protected characteristics, and "fair" must be a chosen, tested threshold that someone owns — not an unexamined by-product of the model. Bias that cannot be mitigated to the chosen level is a barrier, not an accepted cost of going live.

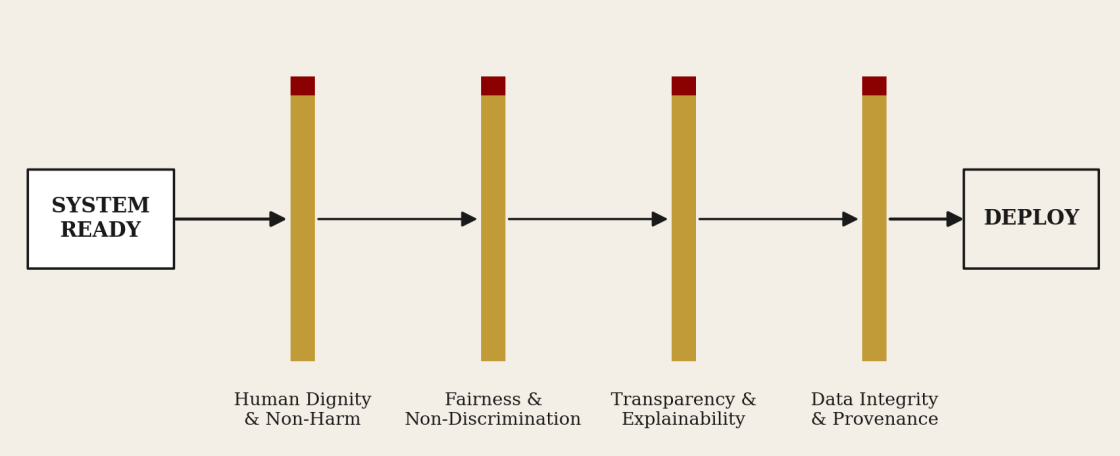

Figure 2 — The four principles as deployment barriers. Each is a point at which a responsible organisation is prepared not to proceed.

Transparency and explainability. A person materially affected by an AI decision has a right to understand the basis of the decision in terms that are meaningful to them. Explainability to an auditor is necessary but not sufficient; a decision explicable only to the people who built the system does not meet the obligation to the person it affected.

Data integrity and provenance. A system may only be trained on and operated with data whose provenance, quality, and consent basis are known, documented, and appropriate to the use. Unknown data is ungoverned data, and a system built on it cannot be responsibly deployed until

its data can be accounted for. Four principles; four points at which a responsible organisation is prepared to stop.

Taken together, the four principles read as a single test applied four times.

Principle	The boundary it sets	What it refuses to be
Human dignity and non-harm	No use that demeans, discriminates against, or causes unjustified harm to individuals or groups	Non-delegable. Not waivable by operational necessity
Fairness and non-discrimination	Outcomes fair across protected characteristics, on a chosen and tested threshold that someone owns	Not an unexamined by-product of the model
Transparency and explainability	A materially affected person can understand the basis of the decision in terms meaningful to them	Not satisfied by explainability to an auditor alone
Data integrity and provenance	Provenance, quality and consent basis known, documented, and appropriate to the use	Not deployable on unknown data, whatever the urgency

Table 1 — The four principles as deployment barriers, and the reclassification each resists.

The third column is where the framework departs from most responsible-AI statements. Each principle is defined as much by what it declines to become as by what it asserts, because every one of them fails in the same way: by being quietly reclassified as a risk to accept.

SECTION SIX

Not a Risk *to Accept.*

The difference between a framework that protects people and one that documents them lies in a single word: whether the four boundaries are treated as barriers or as risks. It is a distinction that sounds procedural and is, in fact, decisive.

The risk-management posture is familiar and, for most exposures, correct: identify the risk, assess its likelihood and impact, decide who owns it, and accept, mitigate, or transfer it. Applied to responsible-use boundaries, however, it quietly converts a limit into a lever. "Manage the bias risk" invites someone, under delivery pressure, to accept it. Once a fairness failure, an unexplainable decision, or an unknown-provenance dataset is something a manager can sign off as an accepted risk, the boundary is no longer a boundary. It is a negotiation, and commercial urgency wins negotiations.

The barrier posture refuses that move. It holds that certain outcomes are not available to be accepted at management's discretion — that where a system would cause unjustified harm, produce unjustifiable disparity, decide in ways it cannot explain to those affected, or run on data no one can account for, the correct answer is that it does not proceed until the condition is resolved. This is not risk-aversion. It is a statement about which trade-offs the organisation has decided, at board level, are not management's to make.

This is why responsible AI cannot be delivered by policy alone. A policy states an intention that commercial pressure can override; a barrier is a constraint that the architecture enforces, whether or not the pressure arrives. The final section is about that architecture — how the four boundaries are made structural rather than declaratory, so that the organisation's intent survives contact with a delivery deadline.

SECTION SEVEN

Governed by *Architecture.*

A boundary that depends on someone choosing to honour it is not a boundary. The Responsible AI Framework therefore operates the four principles as a closed control loop, in which each principle generates obligations that are checked before a system deploys and monitored while it runs — enforced by the same architecture that governs authority, rather than sitting beside it as guidance.

The loop runs from constitution to evidence and back. The four principles are adopted at the board level as constitutional obligations. They descend into policy, then into standards that make each obligation testable — a fairness assessment with a chosen threshold, an explainability standard stratified by audience, a provenance and quality standard for data. Those standards are enforced at deployment gates: no system passes without a current responsible-use assessment, a passed fairness assessment, and certified data provenance. The results are recorded in registers, surfaced to the board through a small set of indicators, and fed back into review. Assessment before deployment; monitoring during operation; evidence throughout.

Crucially, this framework does not float free of the one that governs authority. It attaches to it at specific interfaces. The deployment gate that clears a system for its authority is extended to require a fairness assessment. The register that records a system's authority is extended to carry its responsible-use, bias, and data-provenance status. The people who supervise systems are required to recognise harm and unfairness, not only breaches of limits. The maturity model and the board dashboard are extended to carry responsible-AI signals. The two frameworks are separate obligations sharing one enforcement spine.

The effect is that responsibility ceases to be an assurance and becomes an artefact. When a regulator asks whether a system was fair, the organisation produces the assessment and the threshold that identifies who chose it. When asked whether a person could understand a decision, it produces an explanation that meets the standard the system was built to. When asked what the system was trained on, it produces the provenance record. None of these is a description written after the fact. Each is a precondition the system had to satisfy before it was allowed to act — which is the only form in which responsibility can be shown rather than merely claimed.

Authorisation is also not permanent. The conditions that made a delegation appropriate when it was granted will change, and a record of original authorisation does not, by itself, show that the authority remains appropriate now. A complete architecture therefore tests admissibility continuously (Hossain, 2026), monitoring whether a system's operating conditions still match the

profile under which it was authorised and requiring reassessment when they diverge. The pre-authorisation record answers what was authorised; continuous testing answers whether it still holds. An accountable person needs both answers because a regulator will ask both questions.

SECTION EIGHT

Five Questions *for Your Board.*

The measure of a responsible AI framework is not how comprehensively it articulates its values. It is whether the organisation can produce evidence about the people its AI affects. Each of the following is a question the affected person, their advocate, or a regulator could put to the board — and each should be answerable today, with an artefact rather than an assurance.

Whom could this system harm — and did we check before it went live? For any AI system in operation, can you produce a pre-deployment assessment of who it could harm, across financial, physical, psychological, and social dimensions, and evidence that a material risk of harm was resolved rather than accepted?

What fairness did we choose, and who chose it? Can you state the fairness threshold each consequential system was tested against, show that it was met, and name the person who owns that choice — or was fairness left as an unexamined by-product of the model?

Could the affected person actually understand the decision? For a decision that materially affects someone, can the organisation give that person an explanation they can understand and act on — not only a technical trace that satisfies an auditor?

Do we know what our systems were built on? For every system in operation, are the provenance, quality, and consent basis of its data known and documented — or is any system running on data whose origin no one can account for?

Are there barriers or risks that someone can sign off on? Are the four answers above enforced as deployment barriers by the architecture, or are they risks a manager can accept under delivery pressure? A framework that cannot answer these questions does not protect the people its AI affects. It documents them.

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

REGULATORY AND INSTITUTIONAL SOURCES

ASIC (2024). Report 798: Beware the Gap — Governance Arrangements in the Face of AI Innovation. Australian Securities and Investments Commission, 29 October 2024.

ASIC (2015). Regulatory Guide 271: Internal Dispute Resolution / RG 271 automated decision accountability and explainability obligations.

Office of the Australian Information Commissioner. Privacy Act 1988 (Cth), Australian Privacy Principles 1, 3, 6 and 11 — privacy-by-design and automated decision transparency.

European Union (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), Articles 9, 10, 13 and 14 — risk management, data governance, transparency, and human oversight for high-risk AI systems.

United Nations (2011). Guiding Principles on Business and Human Rights — human rights due diligence.

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). *Continuous Admissibility: A Temporal Governance Framework for Delegated Machine Authority in Regulated Enterprises.* Preprint. Zenodo.
https://doi.org/10.5281/zenodo.20580716

Consulting.AI

Scaling Intelligence

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