A 42 CONSULTING.AI WHITEPAPER

The Line You Don't Cross

Why Certain AI Uses Are Barred Before Any Control Becomes Relevant

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 AI governance begins by asking how to manage a system's risks, and that is a reasonable place to begin: *what could go wrong, and how do we control it?* A great deal of responsible practice is exactly that, done well. But it carries a quiet assumption — that every use is, in principle, available to be managed into acceptability. For most uses, the assumption holds. For some, it does not.

There is a prior question, and it is not about how well a use is controlled. It is about whether the use is one the organisation is prepared to make at all. Some applications of AI are not risks to be mitigated but lines a responsible organisation has decided not to cross, because crossing them cannot be reconciled with treating people as people — whatever controls surround the system. Such use is not redeemed by a strong control environment. It was never on the table.

This paper is about that prior question and about the architecture that answers it. It is the distance between a well-managed system and a use that should never have been proposed — the point at which "we are able to build this" has to meet "we have decided we will not". The paper does not settle on a values statement. It shows where the line sits, how a responsible organisation screens for it before anything else, and why that screen cannot be a risk a manager signs off on under pressure.

It is written for the board and the executive who have been assured their AI is "governed" and want to know whether that governance includes the discipline to refuse. Governance of the ordinary kind measures how well the organisation manages its activities. This is about what it has decided not to do at all — and about making that decision hold on the day a feasible, profitable, and demanded use arrives at the gate.

M Maruf Hossain

SECTION ONE

Executive Summary.

Managing an AI system's risks and deciding whether its use is permitted are two different acts. Most AI governance approaches handle the first and treat the second as a stronger version of it. It is not. This whitepaper establishes three propositions.

First, technical feasibility is not permission. That a system can be built, and built well, settles nothing about whether it should be deployed at all. Whether a use is permitted is prior to, and independent of, how well it is controlled. An organisation that can only ask whether a use is adequately governed has no way to ask whether it is a use it will engage in — and the second question decides the first, not the other way around.

Second, a prohibition is a screen that terminates, not a risk that is weighed. A prohibited use is not a high-rated entry on a register to be mitigated down to tolerance. It is a use the organisation has determined it will not make, so the assessment stops the moment one is identified — before any weighing of controls, benefit, or likelihood. Run a prohibition through risk machinery, and it quietly becomes the most severe kind of risk; and a risk, however severe, is negotiable.

Third, the bar is not management's to move. Whether a use is prohibited is a board-level determination, adopted constitutionally and amendable only by the board that set it. No operational necessity, deadline, or executive judgement can waive it — because a prohibition management can set aside under pressure is a preference, not a prohibition. The point of placing the line above management is to keep it standing when management would find it convenient to move.

This paper sets out the prohibited-use question in full, separates it from the risk management that surrounds it, names the uses that are off the table and the regulation now beginning to name them too, and shows the architecture that enforces the line rather than merely asserting it. It ends with five questions a board should be able to answer about the uses it has refused — not next quarter, but today.

SECTION TWO

Feasible Is *Not Permitted.*

Every proposal to deploy an AI system raises two questions that are easily collapsed into one. The first is whether the use is permitted — whether it is one the organisation is prepared to make at all. The second is whether, assuming it is permitted, the system that performs it is adequately controlled: accurate enough, fair enough, explicable enough, monitored well enough. The second is the familiar territory of risk management. The first comes before it, and it is answered differently.

The two are independent, and the independence runs one way. A use can be flawlessly controlled and still be one the organisation should never make: a system that scores citizens on their private conduct can be accurate, well-documented, and closely monitored, and none of that touches the objection to it. A permitted use, by contrast, can be badly controlled — and the remedy is simply to control it better. Excellence on the second question cannot rescue a failure on the first, because the first is not a question about quality. It is a question about permission.

Figure 1 — The field of feasible uses. Risk management operates inside the permitted region; the prohibited-use screen sits at its edge, asked before any control is assessed.

This is why an organisation can hold an unbroken record of well-managed AI and still deploy something indefensible. Risk management is built to make a use safer; it is not built to ask whether the use should exist. Pointed only inward — at accuracy, fairness, monitoring — it returns a clean report on a system that should never have been proposed, because the objection to that system was never a risk it was scoped to see. The machinery works perfectly and answers the wrong question.

The prohibited-use screen exists to ask the first question, and to ask it first. It is not a more severe risk assessment; it is a different act, applied before risk assessment begins. Its output is binary — permitted or not — and where the answer is not, the process stops. Nothing downstream is reached, because there is nothing to reach: a use the organisation will not make has no control environment to design.

The Responsible AI Framework places this screen at the front of every assessment, ahead of the graded evaluation of harm, fairness, and the rest. The ordering is deliberate. There is no purpose in assessing the impacts of a system that is not permitted to exist, and every purpose in refusing it before effort, sunk cost, and momentum make refusal expensive. A line is cheapest to hold at the start — which is exactly where the architecture puts it.

SECTION THREE

The Uses That Are *Off the Table.*

A prohibition discovered case by case is not a prohibition; it is a series of arguments the organisation must win afresh, under pressure, every time. The Responsible AI Framework therefore names its prohibited uses in advance, as a standing register adopted at board level. Eight categories are established — deliberately broad, so that they capture a principle rather than an exhaustive catalogue of applications — and the register is maintained so that emerging cases falling within a category can be named without reopening the principle itself.

The category names that cannot be reconciled with treating people as people. Several strip away human judgement where it is owed — the automated denial of essential services such as credit, insurance, housing, or medical treatment with no capacity for a person to review and overturn the decision. Several turn the technology against people's own understanding — interactions engineered to deceive people about whether they are dealing with a machine, and systems built to exploit known vulnerabilities of children, the elderly, or people in distress. Others reject the machinery of mass control — surveillance that tracks and profiles individuals at scale without a lawful basis, real-time biometric identification in public spaces, and the social scoring of people based on their conduct or circumstances. Two more guard the ground the framework rests on — models trained on personal data with no adequate consent basis, and emotion recognition used in hiring, education, or law enforcement, where the inference is unreliable, and the stakes are high.

None of these lines waits for a statute to draw it. Each is grounded first in the framework's own principles — human dignity, fairness, transparency, data integrity — and only then in whatever regulation happens to agree. That agreement is now taking shape: the European Union's AI Act, at Article 5, prohibits a converging set of practices, including manipulation that distorts behaviour, the exploitation of vulnerability, social scoring, and untargeted real-time biometric surveillance. But the organisation does not hold these lines because Article 5 names them. It holds them because it judged them incompatible with responsible use, and it would hold them if no regulator had ever spoken. Waiting for the law to name a harm before refusing it is not governance; it is the absence of it.

Because the categories describe principles rather than products, they are read for substance, not for literal fit. A use that falls within the spirit of a category but not its obvious wording is escalated, examined, and — if it belongs there — named within the category rather than argued

around it. And because the list is a board determination, it grows only by board resolution and shrinks by nothing less. What the categories share is their effect on an assessment: where one is met, the assessment does not proceed to weigh it. It ends.

SECTION FOUR

A Screen, *Not a Weighing.*

The difference between a prohibition and a risk is not one of degree, and the architecture depends on never confusing them. A risk is something the organisation accepts, mitigates, or transfers based on a judgement about likelihood and impact. A prohibition is something it does not do. Run the two through the same machinery and the prohibition quietly becomes the most severe kind of risk — and a risk, however severe, remains negotiable. Three differences keep the screen from collapsing into a weighing.

A screen is binary; a weighing is graded. The prohibited-use screen returns one of two answers — the use is clear, or it is prohibited. There is no "high but acceptable", no residual position to sign off, no threshold to be reached with mitigation. The graded assessment that follows deals with Clear, Concern, and Boundary because harm and fairness are matters of degree to be managed. Permission is not a matter of degree, so it is not graded. It is screened.

A screen terminates; a weighing continues. When the screen finds a prohibited use, the assessment stops: it does not pass to the boundary assessment, and there is no outcome to be conditioned or accepted. The finding is recorded, the system owner is told to cease, and the matter is escalated because a prohibited use already in development is a nonconformity, not a design choice. A risk assessment is designed to run until it reaches a disposition; the screen is designed to end the conversation.

A screen precedes; a weighing follows. Order matters as much as outcome. The screen runs before the graded assessment, not beside it, because assessing the impacts of a use that is not permitted is wasted work — and worse, it creates the illusion of a sunk cost that later makes refusal feel unaffordable. Putting the binary question first reserves the expensive machinery of impact assessment for the uses that have earned it: the ones that are permitted and now need to be made safe.

This is why the prohibited-use decision cannot be carried by policy language alone. A policy that says "we do not deploy manipulative systems" states an intention, and under delivery pressure, an intention is precisely what gets reinterpreted. What holds the line is a screen wired into the assessment as a gate — a step that must return clear before anything else proceeds, and whose "not clear" the architecture will not let the organisation route around. The next section shows where that screen sits in the flow; the section after it shows what keeps the screen itself from being waived.

SECTION FIVE

The Assessment That *Holds the Line.*

The prohibited-use screen does not stand alone; it is the second step of a four-stage assessment every AI system must pass before deployment. The stages run in a fixed order, and the order is the argument: the cheap, binary questions come before the detailed ones, so that a use is refused — if it is going to be refused — before the organisation has invested in making it work.

Stage one — intake. A proposed use is submitted and assigned to a qualified assessor before any building work begins. Nothing about the system is assumed; the assessment starts from a description of what the use actually is, not what its proponents would prefer it to be.

Stage two — the prohibited-use screen. The use is tested against the eight categories, read for substance and for reasonably foreseeable misuse rather than stated intent alone. If it is clear, it proceeds. If it is prohibited, the assessment terminates here — the hard stop the whole architecture is built around. A "potential" finding is examined by the accountable officer and either cleared with documented reasoning or confirmed as prohibited.

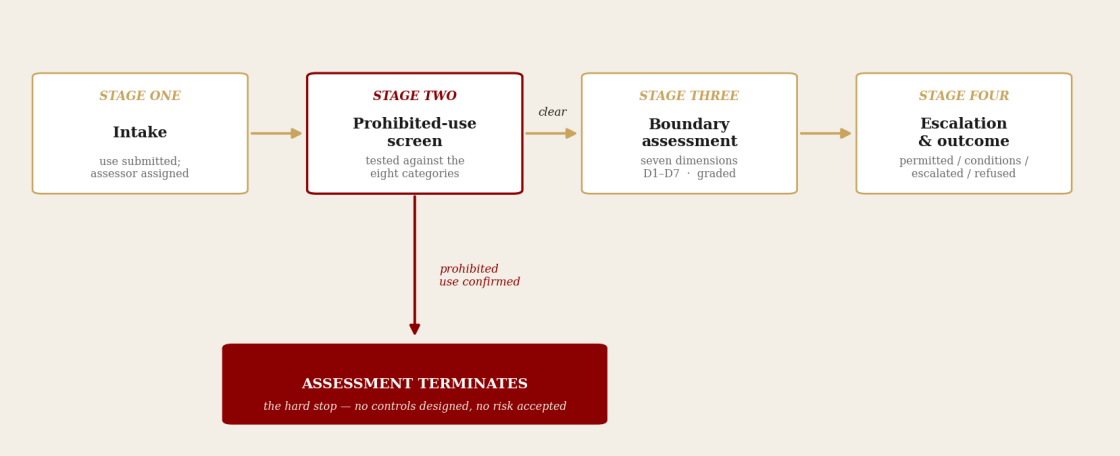

Figure 2 — The four-stage assessment. The prohibited-use screen is a hard stop at stage two; only uses that clear it reach the graded boundary assessment.

Stage three — the boundary assessment. Only permitted uses reach it. Here the system is assessed across seven dimensions — harm, fairness, transparency, data and privacy, dignity and autonomy, systemic impact, and regulatory risk — each rated as Clear, Concern, or Boundary. This is where the degree lives: concerns become documented conditions, and any dimension rated at the boundary triggers escalation. This is risk management, properly placed — after permission has been settled, not instead of settling it.

Stage four — escalation and outcome. Contested and boundary findings follow a defined path: from assessor to accountable officer to risk officer to board committee to full board, and the assessment results in one of four outcomes: permitted, permitted with conditions, escalated, or refused. The screen and the escalation path share one design goal — to ensure that the decision to proceed or not is made at the level with the authority to own it.

SECTION SIX

Not a Risk *to Sign Off.*

A line is only as strong as the organisation's inability to cross it quietly. A prohibited use that a manager can accept as a signed-off risk is not prohibited; it is discouraged. So the final question is not what the lines are but who is able to move them — and the answer has to be structural, because any answer resting on people choosing to hold firm will fail on the day holding firm becomes expensive.

The failure mode is specific and familiar. Under commercial pressure, a prohibition that lives only in policy is reread as guidance, then as a high-rated risk, then as a risk that — given the deadline, the opportunity, the competitor already doing it — someone decides to accept. Each step is locally reasonable. The cumulative effect is that a use the organisation said it would never make is now in production, authorised by no one who believed they were overturning a principle. Nothing was formally waived; the line simply eroded.

This is the structural point developed in the MANDATE preprint series: a constraint that governs only until it is overridden is a policy-level constraint, and a policy-level constraint yields to sufficient pressure (Hossain, 2026). A prohibition has to be of a different kind — one that governs regardless, because it sits above the authority of anyone who might be tempted to move it. The distinction is not rhetorical. It is the difference between a rule that depends on resolve and a rule that needs none.

So the prohibited-use categories are not held in a policy that management approves. They are constitutional — adopted by the full board, placed beyond the reach of any executive or committee decision, and amendable only by the body that set them. The officer who runs the screen cannot waive a finding; escalation runs upward, never around. This is not distrust of management. It is the recognition that the whole value of a line is that it holds when the person at the gate would prefer it did not — and only a line management that cannot move can do that.

SECTION SEVEN

Held by *Architecture.*

A prohibition that depends on someone remembering to apply it is not held; it is hoped for. The Responsible AI Framework therefore does not leave the prohibited-use screen as a document to be consulted. It wires the screen into the same gates and registers that govern everything else, so that a system cannot reach deployment without having passed through it — and cannot later be found to have skipped it.

The line runs from constitution to evidence. The board adopts the prohibited-use categories as constitutional obligations. They descend into the responsible-use policy, and from there into the assessment framework that makes the screen a mandatory, ordered step. No system clears the deployment gate without a completed assessment on record; the screen's outcome is written into the register that tracks the system for its whole life; and that register feeds the small set of indicators the board sees. Refusal, in this design, is not an absence of activity. It is a recorded act, with an owner.

Crucially, the screen attaches to the machinery that already governs the enterprise's AI, rather than sitting beside it as advice. The gate that clears a system for deployment is extended to require that the prohibited-use screen has returned clear. The register that records a system's authority carries its responsible-use status. The people who supervise systems in operation are briefed to recognise a prohibited use should one emerge through drift or feature creep, not only a breach of operating limits. Two questions — is this controlled and is this permitted — are enforced by a single spine.

The effect is that a refusal becomes an artefact rather than an assurance. When a regulator asks whether the organisation considered a prohibited use and declined it, the answer is the assessment, the screen outcome, and the register entry — not a recollection that someone once raised a concern. When asked how the line is kept from moving, the answer is the constitutional adoption and the amendment rule, not a statement of good intentions. A prohibition that can be shown is one that was real; one that can only be claimed was always negotiable.

None of this makes the organisation slower to build. It makes it faster to refuse — the expensive thing to do late and the cheap thing to do early. A framework that can only manage risk subjects, every feasible use for an impact assessment, and discovers objections only at the end, when they cost the most. A framework that screens first spends its effort on the uses worth building and says no to the rest before saying no gets hard. That is what it means for a line to be held by architecture rather than by resolve.

SECTION EIGHT

Five Questions *for Your Board.*

The measure of a prohibited-use regime is not how firmly the organisation states its values. It is whether the board can show that a line was drawn, applied, and held. Each of the following is a question a regulator, a court, or a harmed person could put to the board — and each should be answerable today, with an artefact rather than an assurance.

Do we know which uses we have refused — and did we decide before we were asked? Can you produce a standing register of prohibited AI uses, adopted at board level, that names the lines in advance rather than leaving them to be argued case by case under pressure?

Is the screen the first question, or an afterthought? For every system in operation, was a prohibited-use screen completed before the impact assessment and before deployment — or does the organisation work out how to control a use before establishing that the use is permitted at all?

Is a prohibition a stop, or a risk someone can sign off? When a use meets a prohibited category, does the assessment terminate — or can a manager, under a deadline, accept it as a residual risk and proceed? Can you show a case where the answer was no, and the system did not ship?

Who can move the line? Are the prohibited-use categories constitutional — beyond the reach of executive or committee decision, amendable only by the full board — or do they sit in a policy that management can revise when the line becomes inconvenient?

Can we show the refusal, not merely assert it? If a regulator asked whether you had considered a prohibited use and declined it, could you produce the screen outcome and the register entry — or only the recollection that someone had a concern? A framework that cannot answer these questions does not hold a line. It hopes one holds.

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

European Union (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 5 — prohibited AI practices, including manipulation that distorts behaviour, exploitation of vulnerability, social scoring, and untargeted real-time biometric identification.

ASIC (2015). Regulatory Guide 271: Internal Dispute Resolution — RG 271 obligations for human review of, and accountability for, automated decisions affecting consumers.

Office of the Australian Information Commissioner. Privacy Act 1988 (Cth), Australian Privacy Principles 3 and 6 — collection and use limitations bearing on AI training data, profiling, and biometric identification.

Australian Consumer Law / ASIC Act. ASIC Act 2001 (Cth) s.12DA — prohibition on misleading and deceptive conduct, applicable to undisclosed or deceptive AI interactions.

United Nations (2011). Guiding Principles on Business and Human Rights — human rights due diligence, applicable to mass surveillance and social-scoring uses.

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 Constitutional Prohibited Use Architecture: Why Ethical Boundaries in Enterprise AI Governance Must Be Constitutionally Entrenched, Not Policy-Governed.* Preprint. Zenodo. https://doi.org/10.5281/zenodo.20730346

Consulting.AI

Scaling Intelligence

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