A 42 CONSULTING.AI WHITEPAPER

Explainable to Whom?

Why One Explanation Cannot Serve the Regulator, the Operator, and the Person Affected

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 explainable, and the honest answer is another question: explainable to whom? The word is treated as a single property a system either has or lacks — a box an engineer ticks once. But the explanation that lets a regulator reconstruct and defend a decision is not the account a frontline officer needs to act on, and neither is the plain-language explanation owed to the person the decision was made about. One word, three different obligations.

This matters because the tier that is easiest to satisfy is not the one that protects people. A complete, tamper-evident audit trail satisfies the regulator and the internal auditor — and an organisation can point to it and call the system "explainable" while the person who was refused a loan has been told nothing they can understand or act on. The obligation that is technically hardest and commercially least visible is the one owed to the individual, and it is the one that quietly goes missing.

This paper is about that gap, and about why a single explanation cannot close it. It is the difference between a system that can be audited and a decision a person can understand — and it shows why the individual-facing explanation cannot be bolted on afterwards, but has to be designed into the system before it makes a single decision.

It is written for the board and the executive who have been assured their AI is "explainable" and want to know explainable to whom, in what form, and within what time. Explainability is not one artefact that satisfies everyone. It is three obligations to three audiences, each owed independently — and a framework that meets one while claiming to have met them all has not been transparent. It has been lucky.

M Maruf Hossain

SECTION ONE

Executive Summary.

Explainability is not a single property. It is a set of distinct obligations to distinct audiences, and treating them as one is a governance failure with a name. Most organisations satisfy the audience that is cheapest to satisfy and assume the rest will follow. It does not. This whitepaper establishes three propositions.

First, explainability is audience-relative — "to whom?" is part of the question. The same decision must be explained three ways: to the regulator, as a complete audit trail from which it can be reconstructed and defended; to the operator, as a real-time rationale sufficient to act on or override; and to the affected individual, as a plain-language account of what drove the decision and what they can do about it. No single artefact serves all three, because they serve different purposes for different readers.

Second, the three tiers are independent — none substitutes for another. Satisfying the audit trail does not satisfy the individual, and a plain-language letter that omits the decisive factors does not satisfy the regulator. Each tier must be specified, built, and assessed independently. Designing one artefact to cover all three — or citing one to claim another — is the conflation failure this paper is about.

Third, the individual-facing tier must be designed in, not bolted on. The explanation owed to a person — the principal factors that drove their outcome, and the minimum change that would have altered it — can only be produced by a system built to expose them. It cannot be reverse-engineered from a model that was never designed to answer the question. So explainability is a design and procurement constraint, decided before the system is built, not a report generated after it ships.

This paper sets out the three audiences and their three explanations, shows which tier goes missing and why, specifies what the individual is actually owed, and shows why the capability has to be built in rather than retrofitted. It ends with five questions a board should be able to answer about who can understand its AI decisions — not next quarter, but today.

SECTION TWO

Three Audiences, *Three Explanations.*

Ask what it means for an automated decision to be explainable, and the answer depends entirely on who is asking. A regulator examining a refused loan needs a complete, tamper-evident record from which the decision can be reconstructed and defended months later. A frontline officer acting on the system's recommendation needs, in real time, a clear account of what the system found and how confident it is — enough to agree, question, or override. The person who was refused needs something different again: a plain-language account of what drove the outcome and what, if anything, they can do about it.

These are the three tiers the framework requires; they are not three views of a single explanation. There are three obligations to three audiences, serving three purposes: the audit trail for regulatory defensibility, the operator rationale for human oversight, and the individual explanation for the affected person's rights. Each is precisely defined; each is separately owed; and the artefact that discharges one does not discharge the others (Hossain, 2026).

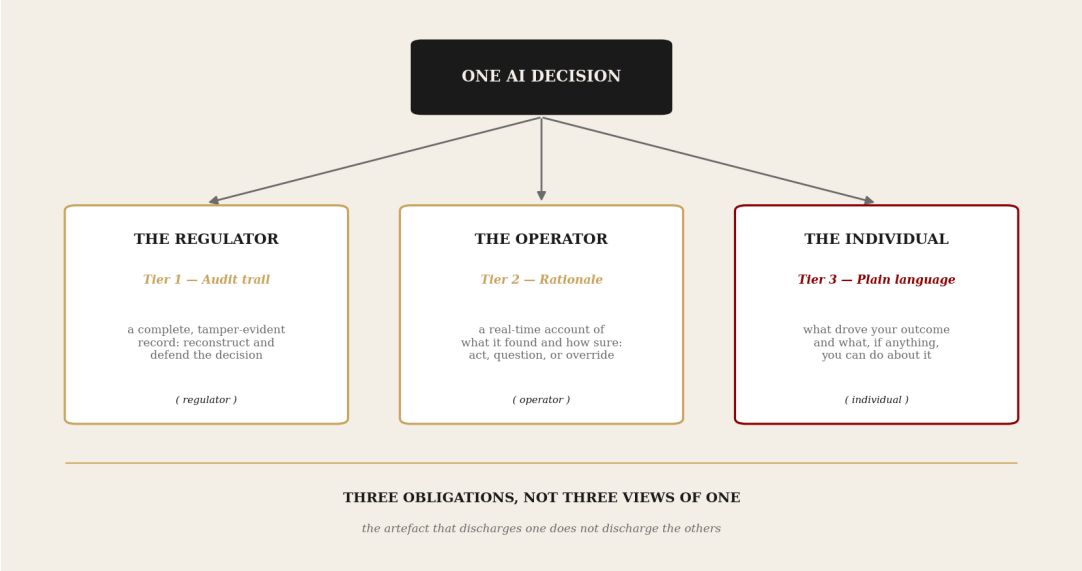

Figure 1 — One decision, three explanations. The audit trail defends the decision to a regulator; the operator's rationale enables human oversight; the individual explanation fulfils the affected person's right to understand. None substitutes for another.

The difference is not one of length or polish — it is a difference of kind. The audit trail is exhaustive and technical, written to survive examination; it records model and prompt versions, inputs, authority scope, and governance state, and no individual is expected to read it. The operator rationale is concise and domain-specific, written for someone with context and a decision to make in the time available. The individual explanation is plain and actionable, written

for someone with no technical knowledge and a stake in the outcome. Optimise any one of them for its audience, and it becomes unsuitable for the others.

So “is it explainable?” is an incomplete and dangerous question because it invites a single yes. The complete question is explainable to whom — and the only honest answer names all three audiences and the distinct artefact each is owed. A system that is fully explainable to a regulator and opaque to the person it judges is not, in any sense that matters, explainable. It is auditable, which is a different and lesser thing.

This is why explainability cannot be treated as one deliverable. The three obligations attach to different readers, carry different content, and are met by different means. The rest of this paper is about the tier most often left unmet — the one owed to the individual — and why meeting it must be a decision made before the system is built, not a report requested after it has already been decided.

SECTION THREE

The Tier That *Goes Missing.*

Of the three tiers, one is reliably built, and two are reliably at risk — and the pattern is not random. The tier that gets built is the audit trail, because it is the one the organisation needs to defend itself. Logging is a familiar engineering discipline; regulators ask for it directly, and its audience — auditors and examiners — has the standing to insist on it. So, the audit trail tends to exist and is good.

The operator rationale is at moderate risk: it is often approximated by a confidence score or a raw feature list, which is not the same as an account that a non-specialist can act on. But the tier that most reliably goes missing is the individual-facing one — and for a specific reason. Its audience is the person affected by the decision, who has the least power to demand it, the least technical standing to know what they are owed, and no seat at the table where the system was scoped. The obligation is real, but the pressure to meet it is weak.

The result is a system that is "explainable" on paper and unaccountable in practice. The audit trail is complete, the compliance box is ticked, and the organisation can truthfully say it can reconstruct any decision. Meanwhile, the person who was refused receives a generic adverse-action notice that names no actual reason, or a reason so abstract it cannot be acted on. Nothing in the audit trail is wrong; it simply was never meant to be read by them, and it answers a question they never asked.

This is the failure the framework is built to prevent: the quiet substitution of the auditable for the explainable. It refuses to let a complete audit trail stand in for the individual's explanation, because the two serve different people and different ends. Naming the individual-facing tier as a separate, mandatory obligation — with its own content, its own timeframe, and its own accountability — is what stops it from being the tier that always gets dropped when time is short and no one at the table speaks for the person on the other side of the decision.

SECTION FOUR

What the Individual *Is Owed.*

If the individual-facing explanation is a distinct obligation, then it has distinct content — and vagueness is how it fails, even when attempted. A letter that says the decision was made "based on a range of factors" discharges nothing. The framework specifies what a real individual explanation must contain, so that "we explained it" can be tested against what was actually owed. Three elements are mandatory.

The principal factors that drove the outcome. Not the whole model, but the handful of inputs that actually determined this decision — the top few features, in order, with their direction of influence, stated in terms the person can recognise: "your credit history and current debt level were the main factors", not a list of coefficients. This is what makes the explanation true to the decision rather than a plausible-sounding cover story.

The minimum change that would have altered it. For an adverse decision, the person is owed the counterfactual: the smallest change to their circumstances that would have produced a different outcome — "had your income been higher by this much, the decision would have changed". This is what makes the explanation actionable rather than merely informative; it tells the person not only why, but what, if anything, they can do.

A plain-language account, delivered in time. The explanation must be in a language a non-specialist can understand, address the specific decision rather than the system in general, and be delivered within the timeframe prescribed by the obligation — not eventually, on request, or after escalation. An explanation that the person cannot understand, or receives too late to use, has not been delivered. It has merely been generated.

None of these can be produced by a system that was not built to produce them. The principal factors require attribution; the model must be instrumented to compute; the counterfactual requires the system to answer a question about a decision it did not make. That is the connection between what the individual is owed and how the system must be built — and it is the subject of the next section. The content of the explanation dictates the design of the system, not the other way around.

SECTION FIVE

Built In, *Not Bolted On.*

The individual-facing explanation cannot be retrofitted. A model that was not designed to expose why it decided cannot be made to explain itself after the fact by wrapping it in a template — the information a real explanation requires was either captured at the moment of decision or is gone. So explainability is not a reporting feature added at the end; it is a property the system must be built with, which means it is a decision made during design and procurement, before anything is built or bought.

At design — the system must expose its reasons. The capability to identify the principal factors behind each decision and to compute the counterfactual for an adverse one has to be part of the system's architecture: instrumented, tested, and logged at the moment of decision. A model chosen or built without this cannot acquire it later, so the obligation to explain constrains which models and designs are even eligible.

At procurement, the vendor must prove the capability. Where the system is bought rather than built, the same obligation applies and becomes a precondition of purchase: the vendor must demonstrate that the system can produce each required tier before it is approved for deployment. "The vendor's model is a black box" is not an exemption from the individual's right to an explanation; it is a reason not to buy that model for decisions affecting people.

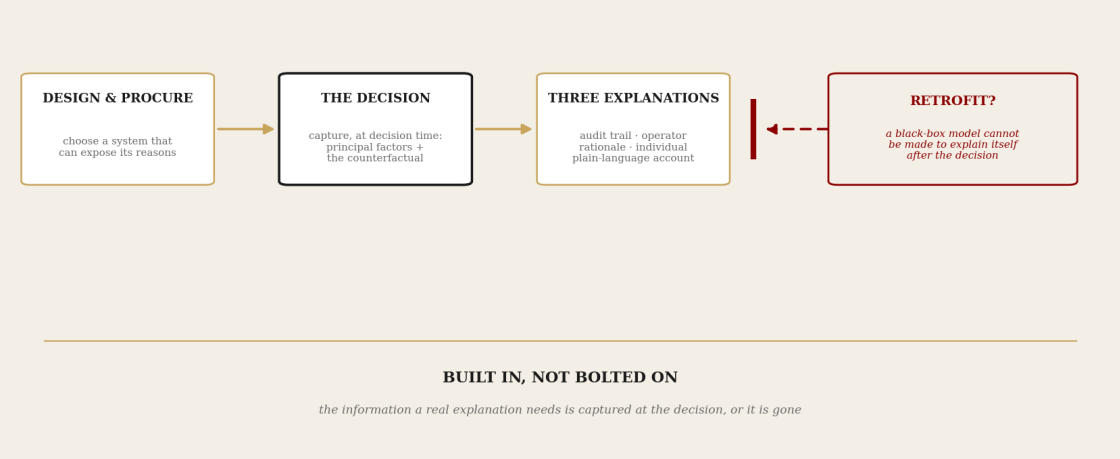

Figure 2 — Explainability as a design constraint. The principal factors and the counterfactual must be captured at the moment of decision; a system not built to expose them cannot explain itself afterwards.

Before deployment, the capability is a gate, not a goal. No system reaches production for individual-affecting decisions without demonstrating every explanation tier it owes. Explainability is verified as a precondition of the deployment gate — not aspired to, not scheduled

for a later release. A system that cannot yet explain itself to the people it will judge is not ready to judge them.

In operation — the capability is maintained, not assumed. Explanation capability degrades if it is not protected: a model update, a feature change, or a pipeline shift can silently break the attribution on which a real explanation depends. So the capability is monitored in production, and a change that breaks it is treated as a fault to be fixed, not an inconvenience to be worked around.

SECTION SIX

One Artefact *Won't Do.*

Everything so far depends on one structural commitment: that the three tiers are independent, and that satisfying one is never evidence of satisfying another. This is the point where explainability regimes most often quietly fail — not by omitting explanation altogether, but by producing one artefact and treating it as though it discharged all three obligations at once.

The failure has a precise shape. An organisation builds a strong audit trail — genuinely good, complete, defensible — and then, under time and cost pressure, treats it as the explanation for everyone. The operator is handed a slice of the log and is expected to make sense of it; the individual is sent a generic notice that alludes to the log's existence. Each tier is nominally "addressed", and none is actually met for its audience. The conflation is rarely deliberate; it is what happens when three distinct obligations are managed as one line item, and the cheapest artefact is quietly asked to cover all of them (Hossain, 2026).

The framework forecloses this by requiring that the tiers be independently specified, built, and assessed. A system's compliance on the audit trail says nothing about its compliance on the individual explanation; each is evaluated on its own terms, against its own audience. "We have a complete audit log" is not an answer to "can the person understand the decision", and the architecture does not let the first stand in for the second. Independence is not bureaucratic duplication; it is the recognition that three different people need three different things.

So explainability is not one deliverable an organisation can produce once and point to. It is three obligations, held apart on purpose, each owed to a reader who cannot be satisfied by the others' artefact. An organisation that has built only the tier it needs to defend itself has met its own interest and left the individual's unmet — and it has done so while able to claim, truthfully but misleadingly, that its systems are "explainable". The whole point of holding the tiers apart is to make that claim testable, one audience at a time.

SECTION SEVEN

Held by *Architecture.*

Three independent obligations, owed to three audiences, are only as real as the organisation's inability to skip the inconvenient one. The framework, therefore, does not leave the individual explanation to good intentions. It wires all three tiers into the same machinery that governs the rest of the AI estate, so that each is specified, evidenced, and checked — and none can quietly go missing.

The line runs from constitution to evidence. The right of an affected individual to understand a decision in terms meaningful to them is adopted at board level as a governing principle; the standard defines the three tiers and what each must contain; the deployment gate requires the tiers a system owes to be demonstrated before it ships; and the register that tracks each system carries its explanation capability and status. Producing an individual explanation and maintaining the capability to produce it are recorded acts with owners — not courtesies extended when someone complains.

Crucially, explainability is attached to the existing enforcement spine rather than sitting alongside it as an aspiration. The gate that clears a system for deployment is extended to require a demonstrated capability to explain for every tier it owns. The register that records a system's authority carries whether it can explain itself to each audience. The people who supervise systems in operation are briefed to notice when the capability breaks, not only when an output is wrong. Explainability becomes a precondition that the system must satisfy and maintain, not a claim in a policy.

The effect is that an explanation becomes an artefact rather than an assurance. When a regulator asks whether the person could understand the decision that affected them, the answer is the delivered explanation, its content, and the timestamp — not a statement that the system "is explainable". When asked which audiences a system can serve, the answer is the register entry, tier by tier. An explainability that can be shown, audience by audience, is one that was actually owed and met; one that can only be asserted was always just the audit trail wearing a broader name.

None of this makes explanation effortless — building a system that can explain why, in terms a person can use, is harder than building one that merely logs what it did. What the architecture provides is not a way around that difficulty but a refusal to let it be evaded: three obligations named separately, built separately, and checked separately, so that meeting the easy one can never again be mistaken for meeting them all. That is the difference between a system that can be audited and a decision a person can actually understand.

SECTION EIGHT

Five Questions *for Your Board.*

The measure of an explainability regime is not whether the organisation can defend its decisions to a regulator. It is whether each audience that a decision touches can actually understand it. Each of the following is a question a regulator, a court, or an affected individual could put to the board — and each should be answerable today, with an artefact rather than an assurance.

Explainable to whom — can we name the audiences? For each consequential system, can you say which of the three explanations it owes — audit trail, operator rationale, individual account — and show that each is actually produced, rather than pointing to one artefact and calling the system "explainable"?

Can the person understand the decision that affected them? For an adverse decision affecting an individual, can you produce the plain-language explanation they received — naming the principal factors and the change that would have altered the outcome — or only the audit log they were never meant to read?

Did we build the capability in, or are we hoping to add it later? Was the ability to explain decisions to individuals a design and procurement requirement, demonstrated before deployment — or is it a retrofit still on the roadmap for systems already deciding people's outcomes?

Do we treat the three tiers as independent? Is each tier specified, built, and assessed on its own — or is a complete audit trail treated as if it discharged the operator's and the individual's explanations, too?

Can we show the explanation, not just assert explainability? If asked whether an affected person could understand a decision, could you produce the explanation delivered to them and its timestamp — or only the claim that the system "is explainable"? A framework that cannot answer these questions has not been transparent. It has been auditable, and I hoped that was enough.

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 13 and 14 — transparency and provision of information to deployers, and human oversight, including explanation of system functioning.

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

Office of the Australian Information Commissioner. Privacy Act 1988 (Cth), Australian Privacy Principle 1 — open and transparent management of personal information, bearing on disclosure of automated decision-making.

International Organisation for Standardization (2023). ISO/IEC 42001:2023 — AI management systems: Clauses 7.5 and 8.6 and Annex A.6.6, documented information and recording of AI system decisions.

United Nations (2011). Guiding Principles on Business and Human Rights — human rights due diligence, applicable to the right of affected individuals to understand decisions made about them.

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). Three-Tier Explainability: A Governance Framework for Audience-Differentiated Explanation Obligations in Regulated AI Systems. Preprint. Zenodo. https://doi.org/10.5281/zenodo.20603034.

Consulting.AI

Scaling Intelligence

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