Why AI Governance Fails at Decision Points, and How Escalation Design Fixes It

*A DEEGS (*Decision Escalation and Exception Governance System) *perspective on authority, escalation, and accountable human judgement*

Shaun Wilde

Core GRC Solutions

Version 1.0 | December 2025

Executive framing

Over the past two years, organisations have invested heavily in AI governance. New standards, regulatory frameworks, and internal policies have emerged at pace. ISO/IEC 42001, the NIST AI Risk Management Framework, and the EU AI Act have all contributed to a clearer shared language for managing AI-related risk. On paper, governance maturity has increased significantly.

Yet accountability failures persist.

When AI systems are involved in material decisions, responsibility is often unclear. Decisions are made quickly, sometimes collaboratively, sometimes implicitly, and often under pressure. When outcomes are challenged, organisations can point to policies, model documentation, and risk assessments, but struggle to explain who was accountable when the decision was made, or why escalation did or did not occur.

This creates a paradox. AI governance frameworks are improving, but confidence in accountable decision-making is not keeping pace.

AI alters decision dynamics in subtle but important ways. It accelerates decision cycles, distributes judgement across people and systems, and introduces persuasive outputs that shape human choice without holding authority themselves. In more advanced uses, AI systems may initiate actions, trigger workflows, or act with delegated authority across multiple systems. Under these conditions, traditional assumptions about review, approval, and escalation begin to break down.

Governance rarely fails because controls are absent. It fails because escalation is undefined, optional, or poorly designed when it is needed most.

Most governance arrangements assume that decisions can be paused, reviewed, or escalated procedurally. In practice, many AI-influenced decisions occur in environments where time pressure, operational complexity, or automation make this unrealistic. Human-in-the-Loop (HITL) controls may exist, but without clear authority boundaries and mandatory escalation triggers, human judgement becomes advisory rather than accountable.

This paper argues that a critical missing control in many AI governance arrangements is explicit decision escalation design. By focusing on authority, escalation, and accountability at decision

points, organisations can move from governance that looks robust on paper to governance that holds under pressure.

What current AI governance frameworks get right

Recent advances in AI governance frameworks represent a genuine step forward. Standards bodies and regulators have moved quickly to establish shared language, expectations, and structures for managing AI-related risk. This progress should not be understated.

ISO/IEC 42001 provides an important management system wrapper for AI. It reinforces governance, accountability, risk management, and continual improvement, and helps organisations integrate AI oversight into existing management systems rather than treating it as a standalone technical concern. In doing so, it correctly positions AI governance as an organisational responsibility, not merely a technical one.

The NIST AI Risk Management Framework (RMF) offers a complementary perspective. By focusing on risk identification, measurement, and mitigation across the AI lifecycle, it encourages systematic thinking about potential harms, controls, and trade-offs. Its emphasis on context, impact, and trustworthiness has broadened discussions beyond model performance alone.

Regulatory approaches, particularly the EU AI Act, further reinforce proportionality through risk classification. By differentiating obligations based on use case, impact, and potential harm, such regimes signal that not all AI systems warrant the same level of control or scrutiny. This aligns with a broader shift toward risk-based governance.

Taken together, these frameworks establish an essential foundation. They clarify responsibilities, raise baseline governance maturity, and provide a common reference point for boards, executives, auditors, and regulators.

However, they also share an implicit assumption: that AI-influenced decisions can be identified, reviewed, and escalated through procedural mechanisms. They largely presume that accountability can be asserted through roles, documentation, and post-hoc review. In environments where decisions are fast, distributed, or partially automated, this assumption becomes increasingly fragile.

In the UK context, this gap is becoming more visible as regulators and auditors place increasing emphasis on accountability, governance outcomes, and evidence of decision-making, rather than reliance on technical controls or documentation alone. Expectations embedded in the UK GDPR accountability principle, sectoral governance regimes, and emerging UK AI policy all reinforce the need to demonstrate who was accountable at decision points, and how escalation operated in practice.

This is not a flaw in the frameworks themselves. It is a gap between governance as described and governance as experienced at decision points. It is within this gap that accountability most often erodes in practice.

Where AI governance breaks down in practice

In practice, AI governance tends to break down not at the level of policy or intent, but at the point where decisions are made under real conditions. These failures rarely appear dramatic at first. They emerge quietly, through ambiguity, speed, and diffuse responsibility.

A common failure mode is optional escalation. Many organisations define escalation paths but treat them as guidance rather than obligation. When AI systems contribute to decisions, particularly through recommendations or risk scoring, there is often no clear threshold at which escalation becomes mandatory. Individuals are left to judge whether a situation merits escalation while balancing delivery pressure, confidence in the system, and perceived expectations.

A second failure mode is implicit authority. AI outputs frequently carry persuasive weight, especially when they appear objective, consistent, or data driven. Over time, recommendations may be treated as defaults rather than inputs. Humans remain nominally accountable, but authority has shifted in practice. Responsibility is retained without corresponding control.

Human-in-the-Loop controls are particularly vulnerable in this context. In many organisations, HITL is implemented as a review step rather than a governance control. Humans are asked to approve or override AI outputs without clear authority boundaries, escalation rights, or protection when decisions are contested. Under pressure, this can lead to passive approval or reliance on the system to justify action.

As AI systems become more interconnected, these issues compound. When AI-driven insights trigger downstream actions or automated responses, decisions are no longer discrete events. Authority becomes distributed across systems, teams, and time. It is often unclear where a decision occurred, or who had the right and obligation to escalate.

Across these scenarios, the pattern is consistent. Governance assumes rational, linear decision-making. Practice is shaped by speed, uncertainty, and delegation. When outcomes are challenged, organisations can demonstrate compliance with frameworks, but struggle to evidence that escalation occurred when it should have, or that a named individual held authority to intervene.

This is not a failure of ethics or intent. It is a failure of design.

Decision escalation as a governance control (DEEGS)

If AI governance is to hold under real operational conditions, escalation cannot remain a procedural afterthought. It must be treated as a governance control, designed explicitly around decision authority rather than process compliance. This is the role of a Decision Escalation and Exception Governance System (DEEGS).

DEEGS is not a replacement for existing frameworks, nor an additional layer of policy. It is a governance outcome focused on how authority, accountability, and escalation are enforced when decisions are made. Where traditional frameworks describe what should exist, DEEGS is concerned with what must happen when decisions cross defined boundaries.

At the core of DEEGS is the recognition that not all decisions are equal. Decisions differ in impact, reversibility, novelty, and risk. In AI-influenced environments, these characteristics can change dynamically, particularly where systems operate at speed or exercise delegated authority. DEEGS therefore begins by defining decision classes and the authority associated with them.

Escalation in a DEEGS model is mandatory, not discretionary. Escalation triggers are defined in advance and linked to decision characteristics rather than individual judgement alone. When thresholds are crossed, escalation occurs by design, regardless of time pressure or operational convenience.

This reframes HITL from a review activity to a governance control. Humans are not asked merely to approve or override AI outputs. They are assigned explicit authority, escalation rights, and accountability for defined classes of decision. Importantly, escalation is treated as evidence that governance is functioning, not as a failure of performance.

Escalation without authority is ineffective. DEEGS therefore requires that individuals receiving escalations have the right to pause, redirect, or reverse decisions. Override and rollback rights are integral, particularly where AI systems initiate actions or trigger downstream processes. Reversibility and kill-switch mechanisms become part of governance design rather than technical contingency.

Finally, DEEGS treats decision logging as a system of record rather than an audit artefact. The purpose of logging is to evidence that authority was correctly exercised and escalation occurred when required. This produces assurance grounded in decision discipline rather than retrospective justification.

What “good” looks like under pressure

Effective AI governance is not demonstrated by the absence of incidents, but by how decisions are handled when conditions are uncertain, time-constrained, or contested. Under pressure, well-designed governance arrangements display consistent, observable characteristics.

Decisions have clear ownership. For each decision class, a named individual holds authority and accountability. Advice may be collective, and inputs may be automated, but accountability is explicit at the point of decision.

Escalation thresholds are defined in advance and enforced consistently. Triggers are linked to decision characteristics such as impact, uncertainty, novelty, reversibility, or regulatory exposure. When thresholds are crossed, escalation occurs as a matter of design, not discretion.

Authority boundaries are explicit. AI systems, operational teams, and decision-makers operate within clearly defined limits. Delegated or agentic behaviour is treated as an exception requiring heightened governance, not as a default mode of operation.

Escalation is supported by real authority. Individuals receiving escalations have the practical ability to pause actions, redirect outcomes, or reverse decisions. Override and rollback mechanisms are integral components of governance design.

Exceptions are expected and designed for. Effective governance anticipates edge cases, model drift, and unexpected interactions. Exceptions are surfaced, escalated, and owned rather than suppressed or rationalised.

Finally, governance produces evidence that withstands scrutiny. Decision records demonstrate not only what was decided, but that authority was exercised appropriately and escalation occurred when required. This supports executive assurance, board oversight, audit scrutiny, and regulatory engagement without reliance on reconstruction after the fact.

Implications for boards, executives, and auditors

The increasing role of AI in organisational decision-making has material implications for governance and assurance. Across boards, executives, and auditors, confidence in AI depends on confidence in how decisions are escalated and owned under pressure.

For boards, assurance must extend beyond the existence of frameworks and policies. Boards should expect evidence that decision authority and escalation are explicitly designed, enforced, and tested. Questions of AI governance are questions of accountability at decision points, not abstract technical compliance.

For executives, governance must protect rather than diffuse responsibility. As AI accelerates and distributes decision-making, executives remain accountable for outcomes even where authority has drifted. Explicit escalation design ensures that difficult decisions are surfaced, shared, and evidenced rather than silently absorbed.

For auditors and regulators, scrutiny will increasingly focus on decision evidence rather than intent. The ability to demonstrate that escalation occurred when required, and that accountable humans intervened appropriately, will become central to credible assurance.

Across all roles, the conclusion is consistent. AI governance cannot be assessed solely through static artefacts or model-level controls. It must be evaluated through the discipline of decision-making itself. Organisations that design escalation structurally will be better positioned to demonstrate accountability, withstand scrutiny, and maintain trust as AI becomes embedded in real operational decisions.

References and governance anchors

International standards and frameworks

- ISO/IEC 42001:2023, *Artificial intelligence management system*
- NIST, *AI Risk Management Framework (AI RMF 1.0)*, 2023
- European Union, *Artificial Intelligence Act*

UK governance and regulatory context

- UK Information Commissioner's Office (ICO), *Accountability Framework*
- UK Government, *A pro-innovation approach to AI regulation* (AI Regulation White Paper)
- UK Government, *National AI Strategy*