THE INVISIBLE EVALUATOR

How AI Is Changing the Way Institutions Judge, and What Governance Must Do About It

---

A research paper on AI governance in institutional evaluation

Priscilla Osaro | Inside the Bid Room

Abstract

Across procurement, grant-making, hiring, and academic review, institutional evaluation is undergoing a fundamental transformation. The judgment of qualified human evaluators is increasingly being shaped, filtered, and framed by AI systems before a human reviewer forms an independent opinion. While these systems are typically introduced as productivity tools to manage growing volumes of information, they are increasingly influencing how institutional judgments are made in ways that often remain invisible to governance processes.

This paper examines that shift through one of its clearest contemporary testbeds competitive proposal and bid evaluation, where AI is used screen, summarize, score, rank, and assess submissions under conditions of information asymmetry, high stakes, and evolving institutional oversight. It introduces the concept of the **Invisible Evaluator** to describe AI systems that materially influence institutional evaluation without themselves being recognized, governed, or held accountable as evaluators.

Drawing on recurring patterns in donor-funded procurement, grant review, and government contracting, the paper argues that the principal governance challenge is not that AI-assisted evaluation will produce obviously incorrect decisions, but that it can quietly reshape the criteria, framing, and exercise of human judgment in ways that institutions neither monitor nor systematically govern. It identifies four interlocking governance risks, automation bias, hidden criteria, diminished explainability, and diffused accountability and proposes a practical governance framework built around four institutional commitments: visibility, calibration, contestability, and human ownership of judgment.

The paper concludes with recommendations for organizations that build, procure, or deploy evaluative AI, arguing that the responsible adoption of AI in institutional decision-making depends not only on improving efficiency, consistency, or throughput, but on ensuring that the governance of human judgment evolves alongside the technologies designed to support it.

1. Introduction: The Rise of the Invisible Evaluator

Every institution that allocates something scarce, a contract, a grant, a job offer, or a place in a program depends on evaluation. Someone, or some process, must assess competing claims and decide which one succeeds. For most of institutional history, that someone has been a person: a procurement officer, a grant panel, a hiring committee, or a peer reviewer. Their judgment was imperfect, sometimes biased, and occasionally shaped by politics, fatigue, or institutional constraints but it was visible. It could be questioned, appealed, audited, and, over time, improved.

That is changing, and it is changing quietly. AI systems are increasingly being used in the evaluation stages of high stakes decisions: screening applications before a human reviewer sees them, scoring proposals against evaluation rubrics, summarizing lengthy submissions into digestible briefs, and flagging which bids or applications merit closer attention. In most cases, no single decision was made to hand judgment to AI. Instead, AI entered as a productivity tool a way to manage growing volumes of information, reduce administrative burden, and improve consistency and, over time, elements of evaluative judgment migrated to it gradually, incrementally, and often without explicit institutional recognition.

This paper introduces the concept of the **Invisible Evaluator**: an AI system that materially shapes which options are seen as viable, how they are interpreted, and how they are scored, without itself being recognized, governed, or held to the standards institutions apply to human evaluators. The Invisible Evaluator does not emerge because institutions consciously decide to replace human judgment. Rather, it develops incrementally as AI tools introduced to support evaluation begin to influence the judgments they were intended only to assist. The result is not the disappearance of the human evaluator, but a subtle redistribution of judgment within the evaluation process itself.

Existing discussions of AI governance have rightly focused on issues such as fairness, bias, transparency, explainability, and accountability (Mittelstadt et al, 2016 NIST AI RMF, 2023 OECD, 2019). Comparatively less attention has been given to how AI becomes embedded within institutional evaluation not as the final decision maker, but as a system that increasingly shapes what human evaluators see, prioritize, and ultimately judge. This paper addresses that gap by examining AI as an institutional evaluator rather than simply another decision-support technology, arguing that its governance requires closer attention than current oversight frameworks typically provide.

The stakes of getting this wrong are not abstract. Evaluation determines who receives funding, who gets hired, whose infrastructure gets built, whose research is supported, and whose ideas are taken seriously. When the mechanisms shaping those decisions become invisible, three governance failures begin to emerge. First, the criteria that influence outcomes gradually diverge from the criteria institutions formally state. Second, responsibility for mistakes becomes increasingly difficult to locate as judgment is distributed across people, processes, and AI systems. Third, the individuals best positioned to detect when something has gone wrong experienced human evaluators are progressively removed from the earliest and often most consequential stages of evaluation.

This paper is written for those who operate within that emerging governance challenge: leaders in AI governance, responsible AI practitioners, procurement and grant-making institutions, public-sector organizations, and the developers and deployers of evaluative AI systems. It does not argue that AI should be excluded from institutional evaluation. The efficiency gains AI offers in managing volume, improving consistency, and supporting human reviewers are both real and valuable. Rather, it argues that institutional evaluation is fundamentally an exercise of judgment, and judgment cannot be delegated, distributed, or substantially shaped by AI without expanding the governance responsibilities of the institutions that choose to deploy it.

A Note on Method

This paper is primarily a conceptual policy analysis. The governance risks discussed in Section 4 draw on established research in human automation interaction, judgment and decision-making, and algorithmic accountability, cited throughout and listed in the references. The procurement case study in Section 3 draws on the author's professional experience in donor-funded procurement, competitive proposal evaluation, and government contracting. These examples are presented as illustrative patterns that demonstrate how established governance risks manifest in institutional evaluation, rather than as findings from a formal empirical study. The paper's principal contribution is conceptual: it introduces the **Invisible Evaluator** as a governance lens for understanding AI-assisted institutional evaluation and proposes a practical framework for governing its use.

Limitations. This paper is conceptual it synthesizes existing literature and professional experience to develop a governance framework for AI-assisted institutional evaluation. While procurement serves as the principal case study, the broader applicability of the proposed framework should be tested through future empirical research across additional institutional settings.

2. How AI Is Entering Evaluation Workflows

AI rarely enters an evaluation process as "the evaluator" More commonly, it is introduced to support tasks adjacent to evaluation, and over time the boundary between supporting the evaluator and influencing the evaluation itself begins to erode. Rather than replacing human judgment outright, AI becomes embedded across multiple stages of the evaluation workflow, with each stage appearing operationally modest in isolation but collectively reshaping how institutional judgments are formed. Several entry points recur across sectors.

Screening and triage. When the volume of applications, bids, or submissions exceeds what evaluation teams can reasonably review, AI is increasingly used to produce an initial shortlist. This first stage determines which submissions receive meaningful human attention and which do not. Because submissions excluded at triage often receive no further review, the triage stage can become one of the most consequential points in the evaluation process, even when the exclusion itself is generated by an AI-assisted workflow rather than a direct human decision.

Summarization. Lengthy technical proposals, grant applications, or research submissions are increasingly condensed into concise briefs for time constrained reviewers. In practice, the summary, not the original submission becomes the primary object of evaluation. Whatever the summarization system emphasizes, downplays, or omits shapes the information available to the reviewer and, consequently, the judgment that follows.

Scoring against rubrics. AI tools are increasingly used to generate preliminary scores against published evaluation criteria, leaving human reviewers to accept, modify, or override those assessments. Research in decision science has consistently demonstrated the influence of anchoring effects: an initial score or recommendation can shift the range that reviewers subsequently regard as reasonable, even when they believe they are exercising independent judgment. As a result, preliminary AI-generated assessments may influence final decisions even when humans retain formal decision-making authority.

Comparative ranking. Some AI systems extend beyond scoring individual submissions by ranking them relative to one another before a review panel convenes. In these cases, AI does not simply evaluate individual applications; it begins shaping the competitive landscape itself by influencing which submissions appear strongest before independent deliberation has begun.

Compliance and red-flagging. AI is also used to assess compliance with formal requirements such as page limits, mandatory certifications, eligibility criteria, or budget formatting. At first glance, these appear to be administrative rather than evaluative tasks. In practice, however, compliance flags frequently function as de facto exclusion mechanisms, particularly where reviewers rely on automated checks to determine whether submissions proceed to substantive evaluation.

What unites these entry points is that each appears, on its own, to be a reasonable efficiency measure. None represents an explicit institutional decision to replace human evaluators. The governance challenge emerges from their cumulative effect. Institutions rarely examine what happens when AI-assisted screening, summarization, scoring, ranking, and compliance assessment are layered together within a single evaluation process. By the time a proposal, application, or submission reaches a human decision-maker, it may already have passed through several AI-mediated stages, each shaping the information, options, and judgments available to the reviewer.

From the evaluator's perspective, the final decision may still feel entirely human. From the perspective of the evaluation system, however, that decision has already been influenced by a sequence of AI-assisted judgments that determined what the evaluator saw, how information was presented, and which options appeared worthy of consideration. It is this gradual redistribution of evaluative influence not the replacement of human decision-makers that gives rise to the Invisible Evaluator.

3. Case Study: Procurement and Proposal Evaluation

Competitive proposal evaluation—whether in donor-funded development tenders, government contracting, or corporate procurement—provides an especially revealing case study for understanding the governance implications of AI-assisted evaluation. It combines many of the conditions under which evaluative AI is both most attractive and most consequential: high application volumes, significant financial and social stakes, compressed evaluation timelines, and reviewers working under sustained institutional pressure to make consistent decisions efficiently. While the examples that follow are drawn from procurement, the governance dynamics they illustrate extend well beyond it.

Consider a typical donor-funded procurement process. A call for proposals may attract dozens or even hundreds of submissions, many offering broadly similar approaches and using comparable technical language, as proposal writing has itself evolved into a specialized

discipline with recognizable conventions, structures, and templates. Evaluation panels composed of volunteers, technical experts, or short-term contracted reviewers must assess substantial technical and financial documentation within limited timeframes. Under these conditions, AI-assisted screening, summarization, and preliminary scoring naturally emerge as attractive solutions to an existing operational bottleneck.

Within this environment, at least two recurring governance risks become apparent, independent of any particular AI model or software vendor.

The recycled proposal problem. Many organizations reuse successful proposal language, structures, and even substantial sections from previous submissions, whether their own or drawn from publicly available award documentation. If AI-assisted evaluation systems are trained, prompted, or calibrated to recognize patterns associated with historically successful proposals, they may begin rewarding familiarity rather than merit. Over time, proposal characteristics correlated with previous success become proxies for quality. The result is an evaluation process that increasingly favors proposal writing sophistication, institutional memory, and established organizational capacity over the underlying quality, feasibility, or contextual appropriateness of the proposed solution. A well resourced repeat applicant may therefore outperform a first time applicant with a stronger locally grounded intervention, not because the proposal is objectively better, but because both AI systems and human reviewers become anchored to patterns that resemble previous winners.

The budget-narrative mismatch problem. Proposal evaluation also requires determining if technical activities and financial allocations genuinely support one another. This demands evaluative judgment than simple consistency checking. AI systems can efficiently verify that activities described in the technical narrative correspond to budget lines and that required documentation appears internally consistent. They are considerably less equipped to assess if those activities are appropriate, realistically costed, or proportionate to the proposed outcomes. A proposal may therefore receive a positive consistency assessment while still containing unrealistic assumptions, weak implementation logic, or poorly justified resource allocations. Under time pressure, reviewers may unintentionally interpret an AI-generated indication of consistency as evidence that broader evaluative questions have already been examined.

these examples share a common structure. The AI system is not malfunctioning. It is performing precisely the task it was designed to perform. The governance risk arises when the narrow competence of that task is interpreted by human reviewers as a broader endorsement of proposal quality than the system was ever designed to provide. Gradually, the criteria the AI system optimizes for pattern familiarity, structural consistency, formatting compliance, or historical similarity begin to substitute for the criteria the institution intended to evaluate, including substantive merit, contextual relevance, feasibility, innovation, and value for money.

Although illustrated through procurement, this governance pattern is not unique to proposal evaluation. Grant review panels, academic peer review, scholarship selection, vendor assessments, and hiring processes all confront similar conditions: limited human attention, increasing application volumes, and institutional incentives to improve efficiency through automation. Across each of these settings, AI enters to solve a workload problem but gradually begins reshaping the standards by which success is recognized.

A third governance pattern concerns applicant behavior rather than the AI system itself. Once applicants believe that AI participates in screening, scoring, or ranking submissions, they begin

optimizing their documents to align with the perceived preferences of AI-assisted evaluation, with the institution's underlying objectives becoming secondary. Language becomes increasingly standardized, documents become more keyword oriented, and proposal structures converge toward whatever appears most likely to satisfy automated evaluation. This mirrors the evolution of search engine optimization after search algorithms became the primary gateway to online visibility. Such adaptation is entirely rational. Yet it subtly changes what the institution is rewarding. Success increasingly reflects the applicant's ability to anticipate AI-mediated evaluation rather than their ability to demonstrate substantive capability or deliver meaningful impact.

The consequences are particularly significant for under resourced applicants. Smaller organizations, first-time bidders, locally led institutions, and organizations without dedicated proposal development capacity may face a double disadvantage: fewer resources to produce polished AI-optimized submissions and less understanding of how AI-mediated evaluation influences selection outcomes. In donor-funded development especially where institutions frequently seek to broaden participation and diversify funding recipients opaque AI-assisted evaluation can unintentionally undermine those equity objectives, even while improving consistency, speed, and administrative efficiency.

The significance of this case study therefore lies not in procurement itself, but in what procurement makes visible. It demonstrates how AI rarely replaces institutional judgment outright. Instead, it redistributes evaluative influence across multiple small decisions that appear operationally harmless in isolation but collectively reshape how institutions determine merit. Procurement simply provides one of the clearest environments in which that redistribution can be observed.

4. The Governance Risks

The preceding sections illustrate how AI becomes embedded within institutional evaluation. The governance challenge, however, lies not simply in AI's presence but in the specific risks that emerge once evaluative influence is distributed across both human and AI-assisted processes. Four governance risks recur across these settings. They are closely interconnected, and addressing any one of them in isolation is unlikely to produce meaningful oversight because each reinforces the others.

4.1 Automation Bias

Automation bias refers to the well documented human tendency to place undue confidence in automated outputs, particularly under conditions of time pressure, uncertainty, or repeated exposure to systems that appear consistently reliable. Within institutional evaluation, this manifests when reviewers treat AI-generated summaries, scores, recommendations, or compliance flags as authoritative starting points rather than as one source of evidence among several.

The governance challenge arises because questioning an AI recommendation often requires reviewers to undertake precisely the work the technology was introduced to reduce: reading original documents, reconstructing evidence, and forming an independent judgment. Under operational pressure, agreement becomes significantly easier than verification.

Automation bias is particularly pronounced where evaluators have limited time, limited subject-matter expertise, or exceptionally high review volumes. Ironically, these are also the environments most likely to adopt AI-assisted evaluation in the first place. Institutions therefore risk creating a feedback loop in which the reviewers least able to detect inappropriate AI influence become those most dependent upon it.

4.2 Hidden Criteria

Institutional evaluations are intended to operate against explicit criteria: published scoring rubrics, terms of reference, eligibility requirements, or evaluation frameworks. AI systems, however, inevitably introduce a second layer of implicit criteria derived from model architecture, training data, prompting strategies, optimization objectives, or implementation choices. These implicit criteria are rarely identical to those formally adopted by the institution and may not be fully understood even by those deploying the technology.

This dynamic helps explain patterns such as the recycled proposal problem discussed earlier. If an AI system implicitly rewards characteristics associated with historically successful submissions, those characteristics gradually become proxies for quality, regardless of whether they correspond to the institution's stated objectives. The result is not necessarily inaccurate evaluation but a subtle redistribution of evaluative weight away from published criteria toward criteria embedded within the AI system itself.

Hidden criteria are particularly difficult to govern because they rarely appear as obvious errors. Rather than producing incorrect decisions, they quietly reshape what the institution is actually measuring without requiring any formal change to its evaluation framework.

4.3 Explainability

Organizations seeking to audit AI-assisted evaluation frequently encounter a practical limitation: many AI systems cannot explain their outputs to a standard proportionate to the significance of the decisions they influence. A summary may omit important information, or a preliminary score may assign disproportionate weight to one evaluation criterion, without producing an explanation that allows reviewers, auditors, or applicants to understand why.

This creates a fundamental governance challenge. When an unsuccessful applicant, an internal reviewer, or a regulator asks why a submission received a particular outcome, institutions may be unable to provide a meaningful explanation—not because information is being deliberately withheld, but because the AI system itself was never designed to generate explanations capable of supporting institutional accountability.

In evaluation contexts, explainability is therefore more than a technical characteristic of AI systems. It is a prerequisite for appeals, auditing, procedural fairness, and institutional legitimacy.

4.4 Diffused Accountability

Perhaps the most structurally challenging governance risk is diffused accountability. Human evaluation traditionally produces identifiable responsibility. Decisions can be traced to named reviewers, evaluation panels, supervisors, or institutional procedures. AI-assisted evaluation complicates that chain of responsibility.

When an AI-influenced decision produces an undesirable outcome, responsibility becomes distributed across multiple actors: software developers, system vendors, procurement teams, organizational leadership, reviewers who accepted AI-generated recommendations, and policymakers who established the governing framework. Each participant contributes to the outcome, yet no single actor necessarily owns responsibility for the final decision.

This pattern reflects a broader challenge already emerging across AI deployment. Statements such as "the model made an error" risk obscuring the more important governance question: which human decisions enabled that outcome? Decisions to procure the technology, define acceptable oversight, determine reviewer workloads, and establish governance processes remain fundamentally human decisions, even when AI contributes to operational judgment.

Within institutional evaluation, diffused accountability creates an additional governance concern. If no individual or organizational function is clearly responsible for monitoring AI-assisted evaluation, then incentives to identify automation bias, hidden criteria, or explainability failures become correspondingly weaker.

4.5 Why These Risks Reinforce One Another

These four governance risks should not be understood as independent governance failures. They reinforce one another in ways that make isolated interventions insufficient.

Automation bias increases the likelihood that hidden criteria will influence outcomes because reviewers become less likely to question AI-generated recommendations. Hidden criteria become even more difficult to identify when explainability is limited, since reviewers lack meaningful insight into how AI systems reached their outputs. Weak explainability, in turn, makes accountability increasingly difficult because institutions cannot readily determine whether a problematic outcome resulted from model behavior, implementation decisions, reviewer practice, or organizational governance failures. Finally, diffused accountability reduces institutional incentives to improve monitoring, explanation, or independent review, allowing the other risks to persist largely unchecked.

The cumulative effect is greater than the sum of its individual parts. A governance response focused on only one risk, for example, introducing transparency statements without measuring divergence between AI and human judgment, or clarifying accountability without improving explainability may create the appearance of responsible AI governance while leaving the underlying mechanisms of evaluative influence unchanged. The Invisible Evaluator persists not because any single governance safeguard is absent, but because these risks reinforce one another across the evaluation process as a whole.

5. A Governance Framework for Organizations

Addressing the governance risks identified in the preceding section requires more than high-level principles or policy statements asserting that "AI supports rather than replaces human judgment" Such commitments have become increasingly common in organizational AI policies, yet they lack corresponding institutional practices capable of demonstrating whether that principle is being upheld in reality.

The framework proposed in this paper translates the four governance risks into four operational commitments. Together, they provide a practical model for governing AI-assisted institutional evaluation by focusing not only on AI systems themselves, but on the organizational processes through which evaluative judgment is exercised.

Visibility: Make the Invisible Evaluator Visible

Effective governance begins with visibility. Institutions should maintain a comprehensive inventory of every stage within an evaluation workflow where AI influences a submission before a human reviewer forms an independent judgment.

This inventory should function as a living governance document rather than a one-time disclosure. It should be reviewed whenever evaluation workflows change, new AI capabilities are introduced, or existing systems are updated. At a minimum, institutions should be able to answer a straightforward governance question: **Which stages of our evaluation process currently involve AI, and what role does AI perform at each stage?**

Calibration: Measure Divergence, Not Simply Accuracy

Most AI evaluation systems are validated primarily against measures of technical accuracy or consistency. Governance requires a different form of measurement.

Rather than asking only whether AI outputs correspond to historical benchmarks, institutions should regularly measure where AI-assisted judgments diverge from independent human evaluation.

Patterns of divergence deserve as much attention as patterns of accuracy. Persistent divergence within particular applicant groups, proposal types, subject areas, or geographic contexts may indicate hidden criteria emerging within the evaluation process long before those issues become visible through complaints, appeals, or public scrutiny.

Contestability: Enable Review of AI-Influenced Decisions

Institutional appeals mechanisms should extend beyond reviewing final evaluation outcomes. They should also enable meaningful examination of AI-assisted stages that materially influenced those outcomes.

Applicants, reviewers, auditors, or oversight bodies should be able to ask whether AI-assisted screening, scoring, ranking, or compliance assessment affected an evaluation in ways not reflected in published criteria. Contestability therefore requires more than an appeals process; it requires institutional capacity to reconstruct how AI-assisted evaluation contributed to a decision and identify the organizational function responsible for providing that explanation.

Without contestability, procedural fairness becomes difficult to demonstrate because the most influential stages of evaluation remain beyond meaningful review.

Human Ownership of Judgment: Protect the Conditions for Independent Evaluation

Maintaining humans "in the loop" is insufficient if institutional conditions make genuine independent judgment practically impossible.

Organizations frequently introduce AI-generated summaries, scores, or recommendations to reduce reviewer workload. While operationally beneficial, these interventions may also reduce the time, cognitive engagement, and independent analysis required for evaluative judgment. Human ownership therefore depends not merely on preserving a formal human decision-maker, but on preserving the institutional conditions under which independent judgment can still occur.

This includes ensuring that reviewers engage directly with original submissions where appropriate, allocating sufficient review time for meaningful deliberation, and explicitly training evaluators to recognize automation bias and other AI-related governance risks as part of standard evaluation practice.

From Principles to Governance Practice

These four commitments are deliberately structural rather than aspirational. Statements asserting that "AI supports but does not replace human judgment" are difficult to evaluate because they provide no observable governance mechanisms. By contrast, visibility inventories, divergence measurement, contestability procedures, and protected conditions for independent review can all be monitored, audited, and continuously improved. Their value lies not simply in expressing institutional intent but in making responsible AI governance operationally verifiable.

Sequencing Implementation

Organizations begin with contestability because appeals mechanisms are highly visible to external stakeholders and relatively straightforward to communicate in governance documentation. In practice, however, effective implementation depends upon sequencing.

Visibility must come first because institutions cannot govern AI-assisted evaluation that they have not yet mapped. Calibration follows by identifying where AI-assisted judgment diverges

from independent human assessment. Contestability becomes meaningful only once those AI-assisted stages are understood and documented. Human ownership of judgment should be strengthened throughout implementation because it depends primarily on reviewer training, workload design, and organizational incentives rather than technological capability alone.

A practical implementation sequence therefore consists of **Visibility → Calibration → Contestability**, with **Human Ownership of Judgment** operating as a continuous institutional commitment across every stage of deployment.

Assigning Institutional Ownership

Governance mechanisms require accountable owners.

Visibility should ordinarily be owned by the organizational functions responsible for AI procurement, technology governance, or evaluation process design. Calibration should be overseen by monitoring, quality assurance, or evaluation management teams with access to longitudinal performance data. Contestability should be incorporated into existing appeals, grievance, or oversight mechanisms while explicitly extending those functions to AI-assisted evaluation. Human ownership of judgment ultimately rests with the leaders responsible for evaluator capability, workload management, and professional standards.

Without clearly assigned ownership, governance responsibilities become distributed across multiple functions without meaningful accountability. In practice, responsibilities shared by everyone frequently become responsibilities owned by no one.

6. Recommendations for Leaders

The governance framework proposed in this paper establishes the institutional capabilities required to oversee AI-assisted evaluation. The following recommendations translate that framework into practical leadership actions for organizations that develop, procure, deploy, or rely upon evaluative AI systems.

Audit Before You Optimize

Before introducing additional AI capabilities to address workload or efficiency challenges, organizations should first map their existing evaluation workflows. In many institutions, AI has entered evaluation incrementally through separate tools adopted to solve isolated operational problems. As a result, leaders often underestimate the number of AI-assisted interventions that already influence institutional judgment. Governance begins with understanding the evaluation pipeline that already exists before expanding it further.

Treat Evaluative AI as a Governed Category

AI systems that influence hiring, grant allocation, procurement, admissions, or other institutional evaluations should not be governed as ordinary productivity software. Their procurement, approval, documentation, and oversight should reflect the significance of the decisions they influence. The governance requirements applied to evaluative AI should therefore exceed those applied to administrative tools such as scheduling assistants, transcription software, or meeting summarization platforms.

Invest in the Institutional Infrastructure of Oversight

Effective governance depends less on sophisticated AI policies than on the organizational capabilities required to implement them. Divergence monitoring, documentation of AI-assisted workflows, reviewer training, appeals mechanisms, and governance audits may appear operationally routine, yet they determine whether oversight exists in practice rather than only in policy. These functions should be treated as core governance infrastructure rather than discretionary administrative costs.

Preserve Human Expertise as a Governance Resource

Experienced evaluators detect emerging governance problems before formal monitoring systems do. Patterns such as unusually consistent scoring outcomes, recurring omissions in AI-generated summaries, or unexpected proposal rankings are frequently first identified through professional judgment rather than automated monitoring.

Organizations should therefore establish formal mechanisms through which evaluators can report concerns about AI-assisted evaluation and ensure those observations reach the individuals responsible for governance, quality assurance, or system oversight. Preserving experienced human judgment is not simply a staffing issue; it is an essential component of responsible AI governance.

Be Precise About What "Human in the Loop" Means

Public statements asserting that "a human makes the final decision" provide limited assurance unless accompanied by meaningful information about how AI contributes to earlier stages of evaluation. Institutions should communicate clearly which stages of evaluation involve AI, what functions those systems perform, what oversight exists, and where human reviewers exercise independent judgment.

Transparency alone does not eliminate governance risk, but vague assurances make meaningful accountability considerably more difficult.

Evaluate Equity Alongside Accuracy

Organizations should monitor if AI-assisted evaluation disproportionately advantages applicants with greater institutional experience, stronger proposal writing capacity, or greater familiarity with AI-mediated evaluation processes. These questions cannot be answered through technical accuracy metrics alone.

Regular equity assessments should examine whether particular applicant groups, organization types, geographic regions, or first-time participants experience systematically different outcomes after AI-assisted evaluation is introduced. Responsible governance requires monitoring distributional effects as carefully as technical performance.

Review Governance When Systems Change

Governance should evolve whenever evaluation systems evolve. AI models are updated, prompts are revised, evaluation criteria change, and workflows are redesigned far more frequently than most governance reviews occur. Organizations should therefore trigger governance reviews whenever significant changes are made to AI systems, evaluation rubrics, or evaluation workflows, rather than relying solely on annual review cycles.

Aligning governance reviews with operational change helps ensure that oversight remains responsive to the technologies and institutional practices it is intended to govern.

7. Conclusion

The Invisible Evaluator is not a dramatic villain in the story of AI and institutional decision-making. It is something quieter and, in many ways, more difficult to govern: the cumulative effect of reasonable, incremental technology decisions that gradually reshape what institutions recognize as good, credible, or worthy of selection without any single decision-maker consciously intending that outcome. Procurement and proposal evaluation make this pattern particularly visible because the stakes are tangible, evaluation processes are highly structured, and the consequences of subtle shifts in evaluative judgment can already be observed. Yet the phenomenon extends well beyond procurement to any institutional setting in which AI increasingly influences how human judgment is exercised.

This paper has argued that the central governance challenge is not whether AI should participate in institutional evaluation, but how institutions should govern evaluative influence once AI becomes embedded within the decision-making process. In response, it has introduced the concept of the Invisible Evaluator as a way of understanding how AI shapes institutional judgment, together with a practical governance framework built on four institutional commitments: Visibility, Calibration, Contestability, and Human Ownership of Judgment. Together, these provide a foundation for governing AI-assisted evaluation as an institutional responsibility rather than a purely technical capability.

The argument advanced here is not that institutions should remove AI from evaluation. AI offers genuine improvements in efficiency, consistency, and the capacity to manage increasing volumes of information. Rather, the paper argues that evaluative judgment is itself a governed institutional function. When AI materially influences that judgment, governance must evolve alongside it. Responsible AI therefore depends not only on building more capable systems, but also on building institutions capable of understanding, monitoring, and challenging how those systems shape human decision making.