When Humans Must Intervene: A Decision-Grounded Framework for Human Oversight

in Government and Commercial Agentic AI Deployments

Michael Bragen
Principal, ThinkCapital LLC | Director, Government IT and AI Governance Initiative
April 2026

About This Paper

This paper establishes the conceptual and operational foundation for human intervention strategy in agentic AI systems. It identifies five decision characteristics that consistently justify mandatory human review, provides a five-phase implementation framework applicable to both government and commercial settings, and defines the governance structures needed to make oversight durable. It is intended for CIOs, CTOs, AI governance officers, risk managers, and program leads responsible for deploying or overseeing agentic AI at scale.

ABSTRACT

Agentic AI systems are those capable of sequential, autonomous decision-making across extended task chains. Today, these are moving from pilot deployments into operational use in both government and commercial settings. The governance frameworks inherited from earlier, assistive AI have not kept pace. The result is a class of high-consequence decisions being executed without substantive human review, disguised by nominal oversight mechanisms that satisfy policy requirements while providing little actual control.

This paper identifies five decision characteristics that consistently justify mandatory human intervention, regardless of system risk classification, organizational role structure, or audit log architecture: irreversibility, consequence transfer, distributional novelty, value conflict, and legal or regulatory significance. These characteristics function as decision-level intervention standards. A system classified at moderate overall risk can still execute decisions that meet one or more of these criteria and therefore require a human in the execution chain before the action proceeds.

Herein, we provide a five-phase implementation framework for building durable intervention architecture in both government and commercial environments, with particular attention given to the reviewer quality problem, a gap between oversight presence and oversight substance. We propose measurement criteria for distinguishing genuine human control from rubber-stamp compliance. The framework is designed to be operationally deployable without requiring changes to existing AI system architecture, applied at the decision-type level rather than the system level.

EXECUTIVE SUMMARY

The Governance Gap

Agentic AI deployments are not a future scenario. Government agencies and commercial enterprises are operating systems today that execute multi-step autonomous actions. These allocate resources, process applications, route decisions, and manage workflows. Human involvement in many cases is nominal rather than substantive. The oversight mechanisms that exist were designed for assistive AI; systems that generate outputs for human review. They were not designed for systems that act.

The distinction matters operationally. An assistive AI that produces a flawed recommendation can be corrected before harm occurs. An agentic AI that executes a flawed action may not be correctable after the fact. The governance question is not whether humans are formally designated as responsible; it is whether they are positioned to exercise genuine oversight at the moment decisions of consequence are made. Current evidence suggests the answer is frequently no.

Existing frameworks address this problem at the system level, classifying AI deployments by risk tier and applying oversight requirements accordingly. This approach has a structural limitation: risk tier classifications are static, assigned at the system level, and do not account for the heterogeneity of decisions a single system executes. A moderate-risk system can, and routinely does, execute decisions that carry significant, irreversible, or legally accountable consequences. System-level classification provides no mechanism for routing those decisions to human review.

Five Decision Characteristics Requiring Human Intervention

The framework in this paper establishes intervention standards at the decision level rather than the system level. Five characteristics, identified across regulatory frameworks, AI governance

literature, and operational practice, consistently indicate that a human should be in the execution chain before a decision proceeds:

- Irreversibility
- Consequence Transfer
- Distributional Novelty
- Value Conflict
- Legal or Regulatory Significance

See Section 2 for the full decision-characteristic framework.

These characteristics are not mutually exclusive. A single decision can meet multiple criteria simultaneously, and any one criterion is sufficient to require intervention. The framework does not require risk reclassification of the underlying system. It operates as a decision routing rule, applied within existing workflow architecture.

Implementation Framework

Knowing which decisions require intervention is necessary but not sufficient. The paper provides a five-phase implementation framework for building intervention architecture that is both operationally sustainable and governance-durable:

- **Decision Audit.** Map existing agentic workflows to identify all decision types, characterize each against the five criteria, and establish a baseline for current oversight coverage.
- **Intervention Architecture Design.** Define intervention checkpoints at the decision level, specify trigger conditions, establish reviewer roles and authority structures, and integrate them into workflow design before deployment.
- **Reviewer Competency Program.** Establish training and calibration protocols to ensure reviewers at intervention checkpoints are exercising substantive judgment. Competency is the difference between a functioning oversight mechanism and a rubber stamp.
- **Audit Trail Infrastructure.** Implement documentation standards capable of reconstructing the basis for any reviewed decision, including the reviewer's reasoning, not just the outcome. In government settings, this is a legal accountability requirement; in commercial settings, it is a regulatory and litigation risk management requirement.
- **Oversight Quality Measurement.** Deploy metrics that distinguish oversight presence from oversight substance. High approval rates with minimal override frequency are a signal that intervention checkpoints may not be providing genuine control.

The Reviewer Quality Problem

Nominal oversight, defined as human review that exists on paper but provides no genuine control, is the central failure mode in current agentic AI governance. It satisfies policy requirements while eliminating the operational value oversight is supposed to provide. Its presence is more dangerous than its absence, because it creates documented accountability that is not backed by actual human judgment.

The paper addresses the reviewer quality problem directly. Reviewers at intervention checkpoints must have sufficient context to evaluate the decision, sufficient time to apply judgment, defined authority to override or redirect, and feedback on downstream outcomes to calibrate future

reviews. Absent these conditions, intervention checkpoints produce approval artifacts that lack the rigor of oversight. The measurement framework proposed in this paper is designed to detect and surface this condition before it becomes an accountability liability.

Government and Commercial Application

The five-characteristic framework applies across both government and commercial deployments, with context-specific implementation considerations. In government settings, legal or regulatory significance and consequence transfer are frequently the dominant criteria: agencies make decisions that affect citizens and carry statutory accountability requirements regardless of whether AI or humans execute them. The NIST AI Risk Management Framework, OMB M-24-10, and the EU AI Act all articulate human oversight as a requirement; none provides operational criteria for when and how intervention should occur at the decision level. We will explore and expand these criteria below.

In commercial settings, irreversibility and consequence transfer are often the primary drivers: procurement decisions, financial allocations, customer-facing determinations, and supply chain actions that lock in commitments or affect parties who had no role in the decision chain. The legal or regulatory significance criterion also applies with increasing force as AI systems take on compliance-adjacent functions in financial services, healthcare, and other regulated industries.

The full paper develops six operational recommendations covering intervention standards, deployment sequencing, novelty detection infrastructure, reviewer quality, oversight measurement, and documentation requirements. These recommendations appear in **Section 7: Conclusions and Recommendations**, grounded in the implementation framework and governance architecture developed in Sections 3–5.

Intended Audience:
CIOs, CTOs, AI governance officers, risk and compliance managers, program leads, and procurement officers responsible for deploying, overseeing, or contracting for agentic AI systems in government or commercial organizations.

1. The Problem with Current Oversight Models

1.1 From Recommendation to Action

First-generation AI deployments in government were largely advisory. A system would flag an anomaly, suggest a next step, or draft a document. A human would review, decide, and act. The human was always in the execution chain.

That architecture is changing. Agentic AI systems close the loop: they perceive, reason, and act without requiring a human handoff at each step. A system that previously surfaced a recommended contract modification now submits it. A system that flagged a permit application as compliant now approves it. A system that identified a scheduling gap now rebooks the appointment.

This is not a problem with automation, *per se*. The problem is that the oversight models built for advisory systems do not hold when systems become actors. Directing a reviewer to "stay informed" of actions that have already executed is notification, not oversight.

1.2 What Current Frameworks Get Wrong

Most AI governance frameworks in use today rely on one or more approaches to define when humans should be involved:

- **Risk classification:** assign a risk tier to the AI system and require proportional oversight. High-risk systems get more review.
- **Role-based assignment:** designate a responsible official and treat their involvement as satisfying the oversight requirement.
- **Audit and logging:** record what the system did and review it periodically.

Each of these has a structural problem in agentic contexts. Risk classification is a property of the system, not of individual decisions. A system classified as moderate-risk may still execute decisions with irreversible consequences in specific situations. Role-based assignment conflates accountability with oversight: naming a responsible official does not ensure that a human with judgment and context reviews decisions before they execute. Audit and logging are retrospective. They can identify what went wrong but cannot prevent it.

The gap is the absence of a decision-level standard: one that specifies what characteristics of a particular decision, regardless of system risk tier or organizational role structure, require human review before execution.

1.3 The Stakes Are Higher in Government

Government AI deployments carry additional obligations that commercial deployments do not. Decisions affecting citizens must be legally defensible, procedurally consistent, and subject to due process protections. When an automated system denies a benefit, revokes a permit, or initiates an enforcement action, the legal and ethical accountability framework requires that a named human own that decision. Bias, intended or not, undermines the reliability of systems.

The stakes extend beyond individual decisions. Government agencies operate under public trust assumptions that commercial firms do not. An automated decision that is technically correct but perceived as arbitrary or opaque can damage institutional legitimacy in ways that are difficult to recover from. This is an argument for being precise about which decisions require human ownership.

2. Five Characteristics That Justify Mandatory Intervention

The following five characteristics provide a decision-level standard for mandatory human intervention. They are operational criteria derived from where automated systems consistently fail to account for consequences that institutions own.

A decision that meets any one of these criteria should trigger a mandatory human review before execution. A decision that meets multiple criteria should trigger escalated review with documented rationale.

Characteristic	Trigger Condition	Why Automation Falls Short
Irreversibility	Action cannot be undone or is costly to reverse	Automated systems optimize forward and do not price the cost of being wrong
Consequence Transfer	Impact falls on a party outside the decision chain	No mechanism for proxy accountability exists without a named human reviewer
Distributional Novelty	Input is materially outside training or test distribution	The system does not know what it does not know; humans can recognize the signal if it is surfaced
Value Conflict	Decision involves equity, rights, or competing stakeholder interests or biases	Ethical trade-offs are not optimization problems; automation can surface but should not resolve them
Legal/Regulatory Significance	Decision creates obligations, denies rights, or triggers compliance consequences	Accountability requires a named human owner; automated decisions create audit and liability gaps[1]

[1] The legal accountability question for autonomous systems has been stress-tested most visibly in self-driving vehicle litigation. In Uber's 2018 fatal crash in Tempe, Arizona (the first pedestrian death involving an autonomous vehicle) criminal charges were brought against the human safety operator, not the technology company. The operator, not the system, became the accountable party, a pattern legal scholars have called the "moral crumple zone": when an autonomous system causes harm, accountability collapses onto the nearest human in the chain regardless of whether that human had meaningful control. NHTSA investigations into multiple Tesla Autopilot fatalities raised parallel questions about whether human "drivers" were positioned to exercise genuine oversight, or whether system design made nominal oversight structurally impossible. For agentic AI in government and commercial settings, these cases underscore the framework's central argument: naming a human as responsible is not the same as positioning a human to exercise control. Accountability that cannot be exercised is not accountability.

2.1 Irreversibility

The first criterion is the simplest to operationalize and the most commonly overlooked. If an action cannot be undone, or if reversing it is expensive in time, cost, or consequence, a human should approve it before it executes rather than after.

Automated systems are designed to optimize forward. They are not designed to internalize the cost of being wrong. A system that recommends, then executes, a contract termination has no mechanism for pricing in the reputational damage, litigation risk, or operational disruption that follows an incorrect decision. A human reviewer, even a briefly informed one, brings that accounting to the table.

In practice, irreversibility is a spectrum. The threshold question is not whether an action is technically reversible but whether the cost of reversing it is material. A classification change in a case management system that could be corrected in two minutes is not the same as a payment recoupment that triggers a formal administrative process. Organizations implementing this framework should establish explicit thresholds for what counts as practically irreversible in their operational context.

> **Implementation Note: Irreversibility Thresholds**
>
> *Define reversibility in terms of effort, not technical possibility. An action is irreversible from a practical standpoint if correcting it requires a formal process, external notification, or recovery time exceeding a defined threshold (e.g., 4 hours of staff time, external party notification, or regulatory reporting). Document these thresholds in the intervention protocol and review them annually.*

2.2 Consequence Transfer

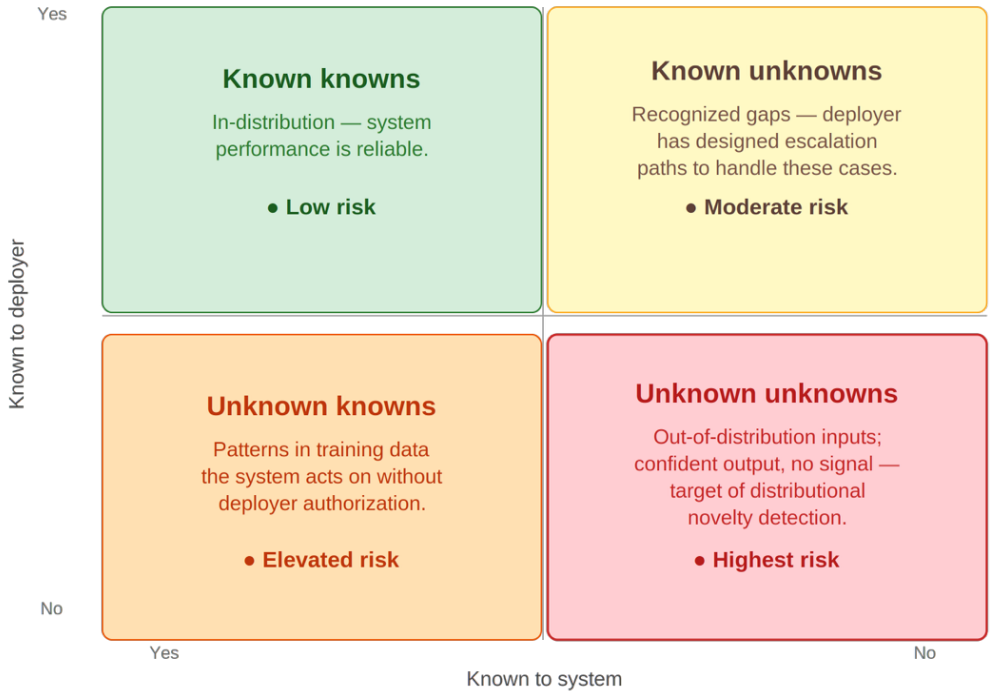
Consequence transfer occurs when the impact of a decision lands on a party who was not part of the decision chain. In government contexts, this describes most decisions that matter: a citizen receiving or being denied a benefit, a permit applicant whose project is approved or rejected, a contractor whose bid is evaluated.

The accountability problem is direct. If an automated system makes a decision and no human reviewed it, who owns the outcome when it is wrong? Audit logs establish what happened. They do not establish that anyone with judgment and accountability was responsible for it happening. Human review is the mechanism that transfers accountability from the system to a person.

This criterion applies with particular force to decisions that affect parties who have no visibility into the decision process. A citizen who receives an automated denial letter has no way of knowing whether a human reviewed their case. Due process and administrative fairness principles in most jurisdictions require that they are entitled to assume one did. Human review is not just good practice in these situations; it may be legally required.

2.3 Distributional Novelty

Distributional novelty occurs when the input to a decision is materially outside the range of cases the system was trained or tested on. This criterion is the most technically complex but also the most important for managing unknown unknowns in deployed systems.[2]



AI systems perform well within their training distribution and degrade in ways that are not always visible at inference time. A system that has processed ten thousand standard permit applications with high accuracy may encounter a combined use application involving a category it has never seen and produce a confident, incorrect recommendation. The system does not know it is out of distribution. It produces an output with the same apparent confidence it would for a standard case.

The human intervention requirement here is conditioned on surfacing. A system can only trigger a distributional novelty checkpoint if it is instrumented to detect and signal when inputs are unusual. This requires investing in confidence calibration, anomaly detection, and what the field calls "I don't know" architecture: the capacity for a system to surface uncertainty rather than produce a confident answer regardless of input quality.

[2] The phrase “unknown unknowns” originates in epistemology and was popularized in policy contexts by Donald Rumsfeld’s 2002 formulation distinguishing what we know we know from what we don’t know we don’t know. For AI risk, a four-quadrant taxonomy is operationally useful. Known knowns are cases within the system’s training distribution where performance is reliable. Known unknowns are recognized gaps or edge cases the system has been explicitly designed to escalate. Unknown unknowns are out-of-distribution cases where the system produces confident outputs without any signal that it is operating beyond its competence. This represents the highest-risk quadrant for agentic deployment. Unknown knowns are patterns embedded in training data that the system acts on without the deploying organization having explicitly authorized them, such as proxy variables for protected characteristics embedded in historical decision records. The distributional novelty criterion in this framework targets the unknown unknowns quadrant specifically: cases where the system does not know what it does not know, and where only instrumented detection can route the case to a human reviewer before execution.

The practical implication is that deploying agentic AI without distributional novelty detection is not a neutral choice. It means that the cases most likely to require human judgment are the least likely to receive it.

> **Implementation Note: Novelty Detection**
>
> *Require that all agentic AI systems deployed in consequential decision contexts include a mechanism for surfacing distributional novelty. This can be implemented as a confidence threshold trigger, an embedding-space distance measure, or a rule-based flag for case attributes outside defined ranges. The signal does not need to be sophisticated; it needs to be present and routed to a reviewer.*

2.4 Value Conflict

Some decisions are not optimization problems. They involve competing interests, ethical trade-offs, or questions about how to weight the claims of different stakeholders when those claims cannot all be satisfied. Automated systems can identify that a conflict exists. They cannot resolve it legitimately.

In government contexts, value conflicts arise wherever equity considerations intersect with eligibility rules, wherever enforcement priorities conflict with community impact, or wherever resource allocation requires trading off one population's interests against another's. These are political and ethical decisions in the true sense: they require someone with democratic accountability or institutional authority to own the choice.

The distinction matters for system design. An AI system that surfaces "this case has characteristics associated with disparate impact" is performing a legitimate and valuable function. The same system making a coverage decision that trades off equity against efficiency is performing a function that belongs to a human decision maker. Designing systems to stay on the right side of this line requires explicit attention during implementation.

Value conflict is also the characteristic most likely to be masked by confident outputs. A system trained on historical decisions that reflected particular value trade-offs will reproduce those trade-offs. Without human review, the choices embedded in the training data become invisible policy.

2.5 Legal or Regulatory Significance

Decisions that create legal obligations, deny rights, or trigger compliance consequences require a named human owner for accountability to function. This criterion operates differently from the others: it is not primarily about the quality of the decision but about the integrity of the accountability chain.

Legal significance in government AI contexts covers a wide range: benefit denials subject to appeal, enforcement actions, contract awards and terminations, regulatory findings, and any decision that creates a record with legal effect. For each of these, administrative and statutory frameworks typically require a human decision maker who can be named, questioned, and held accountable. Automated execution of these decisions does not satisfy those requirements, regardless of the system's accuracy.

In commercial contexts, the equivalent applies to decisions with contractual consequence, consumer protection implications, or regulatory reporting requirements. As AI governance

legislation and rulemaking continue to develop, the range of decisions covered by legal significance criteria will expand. Organizations that build human oversight into these decision types now will be better positioned to demonstrate compliance as requirements mature.

3. Implementation Framework

Establishing the conceptual standard for mandatory intervention is necessary but not sufficient. The framework becomes operational through five implementation phases. These phases apply to both government and commercial settings, with sector-specific notes where the requirements diverge.

Phase	Focus	Key Actions	Success Indicator
1: Classify	Inventory and categorize decision types	Map workflow decisions against the five characteristics; assign intervention tier	Decision taxonomy complete; tiers assigned
2: Design	Build intervention architecture	Define checkpoint triggers, reviewer roles, escalation paths, and timeout rules	Intervention protocols documented and approved
3: Instrument	Implement monitoring and signaling	Deploy confidence thresholds, novelty detectors, and audit logging	System surfaces intervention signals reliably
4: Train	Prepare human reviewers	Build reviewer competency, decision support tools, and calibration exercises	Reviewers can distinguish meaningful signals from noise
5: Measure	Close the feedback loop	Track override rates, intervention latency, and outcome quality; refine thresholds	Continuous improvement cycle operational

3.1 Phase 1: Classify Decision Types

The first task is developing a decision taxonomy for the relevant operational domain. This goes a step further than categorizing the AI system by risk tier. It is mapping the specific decision types the system executes or influences against the five intervention criteria.

For each decision type, the taxonomy should answer: Does this decision type involve irreversible actions? Does it transfer consequences to external parties? Could inputs present distributional novelty that the system would not detect? Does it involve value trade-offs? Does it carry legal or regulatory significance? A decision type that meets one or more criteria is assigned a mandatory intervention tier.

In government contexts, this work should involve legal counsel, program officers, and subject matter experts alongside technical staff. Decision types that appear routine from a technical perspective may carry legal significance or consequence transfer implications that are not visible in the system architecture.

3.2 Phase 2: Design Intervention Architecture

Once decision types are classified, the intervention architecture specifies how and when human review occurs. This design work has four components.

Checkpoint triggers: The conditions under which the system pauses and routes to a human reviewer. Triggers can be automatic (every instance of a high-significance decision type), conditional (when a confidence threshold is not met), or exception-based (when a novelty signal fires).

Reviewer roles and qualifications: Specifying not just that a human must review but what competency that reviewer requires. A legal significance decision requires a reviewer with authority to make that determination. A value conflict decision requires a reviewer with the appropriate programmatic and policy context.

Escalation paths: What happens when a reviewer is unavailable, when a reviewer disagrees with the system's recommendation, or when a decision requires authority above the designated reviewer level. Escalation protocols prevent both bottlenecks and unauthorized delegation.

Timeout rules: Agentic systems operating in time-sensitive contexts need explicit rules for what happens if a reviewer does not respond within a defined window. Default-to-halt is usually the right answer for high-significance decisions; default-to-proceed requires explicit justification and senior approval.

3.3 Phase 3: Instrument for Signaling

Intervention architecture only works if the system reliably signals when intervention is required. Phase 3 implements the technical instrumentation that makes this possible.

Required instrumentation for consequential agentic AI systems includes confidence scoring calibrated to actual decision accuracy, distributional novelty detection that fires before execution rather than after, audit logging that captures not just outcomes but the state of the system at the time of each decision, and routing logic that connects trigger conditions to reviewer queues without requiring manual monitoring.

Government agencies procuring AI systems should require this instrumentation as a contractual deliverable. Systems that cannot surface intervention signals should not be approved for consequential agentic deployment.

3.4 Phase 4: Train Human Reviewers

Mandatory intervention is only as good as the quality of the review it produces. This phase is the most frequently underinvested in AI governance implementations, and the gap shows in practice.

Effective reviewer training addresses three areas. First, decision support: reviewers need to understand what the system is surfacing and what the relevant context is for the decision at hand. Raw model outputs are rarely sufficient. Reviewers need structured presentations of the key factors, the relevant case history, and the basis for the system's recommendation.

Second, signal recognition: reviewers need to understand what a distributional novelty flag means, what a low-confidence score indicates, and how to read the signals the system surfaces rather than defaulting to rubber-stamping the system's recommendation. Automation bias, the tendency to defer to algorithmic outputs even when independent judgment would differ, is a documented and measurable phenomenon that training can reduce.

Third, calibration: reviewer judgment should be periodically assessed against outcomes. When reviewers override system recommendations, what are the outcomes? When they concur, what happens? This feedback loop is the mechanism for continuous improvement of both the system and the reviewers.

3.5 Phase 5: Measure and Refine

Human oversight without measurement is not a quality program. Phase 5 establishes the metrics and feedback loops that make the intervention framework self-improving over time.

Core metrics include the intervention rate by decision type, reviewer override rate by decision type, time-to-review by priority tier, outcome quality for human-reviewed versus system-only decisions, and escalation frequency. These metrics serve two functions: they provide evidence that oversight is substantive (high rubber-stamp rates with no overrides suggest reviewers are not adding value), and they enable threshold refinement (if a decision type triggers intervention rarely and never produces overrides, the trigger may be set too conservatively).

In government settings, measurement outputs should be reported to program leadership and, where legally required, to oversight bodies. The measurement framework is also the primary evidence available when agencies need to demonstrate to auditors, inspectors general, or legislative oversight committees that human oversight of AI systems is substantive.

4. Governance Architecture

A governance architecture for human oversight of agentic AI requires three elements: clear authority structures, defined accountability mechanisms, and institutional resilience over time.

4.1 Authority Structures

Every agentic AI deployment in a consequential decision context should have a designated Human Oversight Lead: a named individual with the authority and responsibility to manage the intervention framework for that system. This role is not the same as the system owner or the program manager, though it may overlap. The Human Oversight Lead is specifically accountable for ensuring that mandatory intervention checkpoints are functioning, that reviewers are trained and qualified, and that measurement data is accurate and acted upon.

Below the Human Oversight Lead, reviewer roles should be defined with explicit qualifications and authority limits. A reviewer who is not qualified to make a legal determination should not be the checkpoint for decisions with legal significance. Authority limits prevent both under-review (routing too high and creating bottlenecks) and over-review (routing too low and producing accountability gaps).

In government settings, the authority structure should be documented in the system's AI impact assessment or equivalent governance artifact and approved at the appropriate senior official level. Changes to the authority structure should trigger a re-approval process.

4.2 Accountability Mechanisms

Accountability requires that someone can be named as responsible for the outcome of a decision and that there is a traceable record connecting that person to the decision they reviewed. In practice, this requires three things.

First, reviewer identity must be logged alongside the decision record. Anonymous review does not satisfy accountability requirements in consequential decision contexts. Second, the reviewer's determination must be explicit: did they concur with the system recommendation, modify it, or override it? A checkbox that records only concurrence does not establish that the reviewer exercised judgment. Third, in decisions with legal significance, the review record should include sufficient documentation to support an appeal or audit, including the basis for the reviewer's determination.

These requirements apply with equal force in commercial contexts wherever decisions have legal or regulatory consequence. As AI governance regulations develop, organizations that have built explicit accountability records will be better positioned to demonstrate compliance.

4.3 Institutional Resilience

Governance frameworks degrade over time if they are not actively maintained. The specific risks in AI oversight governance include reviewer turnover without adequate knowledge transfer, threshold drift as intervention rates decline and oversight becomes nominal, system updates that change decision behavior without corresponding updates to intervention protocols, and scope expansion as systems take on new decision types without re-classification.

Maintaining institutional resilience requires annual reviews of the decision taxonomy and intervention thresholds, mandatory re-onboarding for new reviewers and periodic calibration for existing ones, a change control process that requires governance review for material system updates, and a scope expansion protocol that treats new decision types as requiring fresh classification before deployment.

5. Government-Specific Considerations

While the five-characteristic framework applies in both public and private settings, government deployments carry specific requirements that shape implementation.

5.1 Statutory and Regulatory Constraints

Federal and state administrative law frameworks impose due process, notice, and appeal requirements on agency decisions that affect individual rights and benefits. These requirements were written for human decision makers and do not automatically translate to automated systems.

Human oversight is currently the primary mechanism for ensuring that agentic AI decisions comply with these requirements.

The legal significance criterion in the framework is not a proxy for legal risk: it is a direct reference to the decisions that administrative law reaches. Program managers and AI governance officers should work with agency counsel to identify the specific statutory and regulatory provisions that govern their decision types and ensure that intervention protocols satisfy those requirements.

5.2 Procurement Requirements

Government agencies procuring agentic AI systems should build intervention architecture requirements into solicitations and contracts. Requiring that a vendor provide a system that can execute consequential decisions without intervention signaling capability is equivalent to procuring a vehicle without brakes: technically functional under ideal conditions and dangerous otherwise.

Sample Procurement Language: Intervention Architecture Requirements

The following language is illustrative and should be adapted to agency-specific regulatory and contracting requirements. "The vendor shall provide documented evidence that the system includes: (1) confidence calibration methodology, with calibration curves demonstrating alignment between model confidence scores and empirical accuracy across decision types; (2) distributional novelty detection capability, with configurable sensitivity thresholds and routing logic that triggers human review queues before execution; (3) audit logging architecture sufficient to reconstruct the system state, input data, model output, reviewer identity, and reviewer determination for each decision subject to mandatory intervention; and (4) agency-controlled checkpoint configuration, including the ability to modify trigger thresholds and reviewer routing without vendor involvement. Systems unable to demonstrate these capabilities at contract award shall not be approved for consequential agentic deployment."

Specific procurement requirements should include: documented confidence calibration methodology, distributional novelty detection capability with defined sensitivity parameters, audit logging architecture that supports intervention review records, and configurable checkpoint routing that the agency controls.

5.3 Workforce and Capacity

Government agencies deploying agentic AI in high-volume decision contexts face a specific implementation challenge: if the system is processing thousands of decisions per day, the intervention framework must be designed to handle the corresponding reviewer volume without creating unsustainable workloads.

This requires designing interventions to be genuinely selective, not broadly precautionary. An intervention framework that triggers on a high percentage of decisions may satisfy a checkbox requirement while being operationally unworkable. Selective intervention architecture, supported by good novelty detection and well-calibrated confidence thresholds, allows human review to be concentrated where it matters rather than distributed across the full decision volume.

Implementation Note: Selective Intervention Architecture

Three methods are most commonly used to concentrate human review where it matters. ***Risk-stratified routing*** *divides decision volume into tiers based on the five criteria: decisions meeting multiple criteria go to full review, single-criterion decisions go to expedited review, and decisions meeting no criteria proceed automatically.* ***Confidence threshold triggering*** *routes only decisions below a defined model confidence score, calibrated so that routed volume matches available reviewer capacity without creating backlogs.* ***Attribute-based flagging*** *uses rule-based triggers tied to specific case characteristics, such as claim amounts above a defined threshold, applicant profiles associated with prior override patterns, or decision types with documented distributional novelty. These are designed to catch cases that confidence scoring alone may miss. In high-volume government contexts, all three methods are typically used in combination. Confidence thresholds handle statistical novelty, attribute flags catch known edge-case categories, and risk stratification ensures that decisions meeting multiple intervention criteria receive full review regardless of confidence score.*

6. Commercial Applications

The framework applies in commercial settings wherever agentic AI executes decisions with material consequences for customers, counterparties, employees, or regulated activities.

6.1 Customer-Facing Decisions

Commercial AI deployments that make or strongly influence decisions affecting customers share the consequence transfer and legal significance characteristics with government deployments. Credit determinations, insurance underwriting decisions, employment screening, and content moderation at scale all involve consequence transfer to parties who are external to the decision process. Where these decisions are governed by consumer protection, anti-discrimination, or financial regulation, they also carry legal significance.

Organizations in these spaces should apply the framework with the same rigor as government agencies, even where regulatory requirements do not yet explicitly require it. Regulatory requirements in AI governance are catching up to deployment realities. Organizations that build human oversight into consequential decision processes now will have a compliance and reputational advantage as requirements mature.

6.2 Internal Operational Decisions

High-stakes internal operational decisions, including supply chain commitments, procurement authorizations, and financial transactions above defined thresholds, also warrant framework application. The irreversibility criterion and the value conflict criterion are particularly relevant here. An automated procurement decision that locks in a multi-year supplier relationship is an irreversible decision. An automated resource allocation that systematically disadvantages one business unit involves value conflict even if no external party is affected.

7. Conclusion and Recommendations

The deployment of agentic AI in consequential decision contexts is not a future scenario. It is the current operating environment. The question is not whether AI systems are making decisions that affect people; it is whether the humans nominally responsible for those decisions are positioned to exercise genuine oversight.

The five-characteristic framework in this paper provides a decision-level standard that does not depend on risk tier classifications, organizational role structures, or audit logs to function. It specifies what properties of a decision, in any operational context, require a human in the execution chain before the action executes. The following recommendations translate that standard into operational practice.

7.1 Adopt the Five Characteristics as a Mandatory Intervention Standard

Apply the five characteristics at the decision-type level, not the system level. A moderate-risk AI system can still execute high-significance decisions, and system-level risk classification provides no mechanism for routing those decisions to human review.

Implementation guidelines. Begin by developing a decision taxonomy for each agentic deployment: a structured inventory of every decision type the system executes or materially influences, assessed against all five criteria. The taxonomy should be produced by a cross-functional team that includes legal counsel, program officers, and technical staff, since decision types that appear routine in the system architecture may carry legal significance or consequence transfer implications that are not visible to engineers alone.

Assign each decision type an intervention tier: mandatory review for any type meeting one or more criteria, escalated review with documented rationale for types meeting multiple criteria, and automated processing only for types that meet none. Tier assignments should be treated as governance artifacts, not technical configurations, since they require approval at the appropriate senior level and cannot be changed without a formal review process.

Common implementation failure. Organizations frequently conflate system risk classification with decision-level intervention requirements, applying oversight only to systems classified as high-risk while allowing those systems to execute individual decisions that meet mandatory intervention criteria. The five-characteristic standard corrects this by operating independently of system classification. A procurement system rated moderate-risk that executes contract terminations (and similar irreversible decisions with consequence transfer to external parties) requires human review of those specific decisions regardless of its overall risk tier.

7.2 Invest in Distributional Novelty Detection

A system that cannot surface the signal that a case is unusual cannot route it for human review. Distributional novelty detection is an infrastructure requirement, not an enhancement. Deploying

agentic AI in consequential decision contexts without it is a governance failure regardless of what other oversight mechanisms are in place.

Implementation guidelines. Novelty detection does not require sophisticated machine learning infrastructure. Three approaches are commonly used in combination. Confidence threshold triggering routes decisions to human review when the model's output confidence falls below a calibrated threshold. This should be set so that routed volume is proportional to reviewer capacity, and not so conservative that it triggers on the majority of decisions. Embedding-space distance measurement detects when an input is statistically distant from the system's training distribution, providing a signal that is independent of the model's expressed confidence. Attribute-based flagging uses rule-based triggers tied to case characteristics known to be associated with distributional novelty, such as claim amounts outside historical ranges, geographic or demographic attributes not well-represented in training data, or combinations of case attributes that have not appeared together before.

In procurement requirements, agencies should specify confidence calibration methodology, sensitivity parameters for novelty detection, and the routing logic that connects novelty signals to reviewer queues. Vendors who cannot provide documentation of these capabilities should not be approved for consequential agentic deployment.

What adequate looks like. Systems that produce novelty signals restricted to post-hoc audit logs, or that route flagged cases to a generic review queue with no indication of why they were flagged are inadequate for detecting distributional novelty. A well-instrumented system surfaces a novelty signal before execution, routes the case to a qualified reviewer with a clear indication of why the case was flagged, and logs the reviewer's determination alongside the original signal.

7.3 Design Intervention Architecture Before Deployment

Retrofitting intervention checkpoints into operational agentic systems is significantly more difficult and expensive than building them in. Intervention architecture (checkpoint triggers, reviewer routing, escalation paths, timeout rules, and audit logging) should be delivered at the procurement or development stage.

Implementation guidelines. Intervention architecture design has four components that must be specified before a system goes into production. Checkpoint triggers define the conditions under which the system pauses and routes to a human reviewer: automatic triggers for every instance of a high-significance decision type, conditional triggers when a confidence or novelty threshold is not met, and exception triggers when specific case attributes are present. Reviewer roles and qualifications specify the competency that a reviewer requires. Legal significance decision requires a reviewer with authority to make that determination. Escalation paths must define what happens when a reviewer is unavailable, disagrees with the system recommendation, or lacks authority for the decision at hand. Timeout rules specify what happens when a reviewer does not respond within a defined window: for most high-significance decisions, the correct default is to halt rather than proceed.

Example. A state agency deploying an agentic system to process benefit applications designed its intervention architecture during procurement. Before the system went live, the agency had defined six decision types requiring mandatory human review, specified reviewer qualifications for each, established a 72-hour review window with escalation to a supervisor at 48 hours, and configured the system to halt and notify the applicant if review was not completed within the

window. A different (peer) agency that deployed first and designed oversight later spent eight months and significant staff time retrofitting equivalent controls (during which period the system executed decisions without substantive review.)

7.4 Treat Reviewer Training as a Quality Program

The quality of human oversight is a function of reviewer competency. Calibration and feedback loops are required to maintain it over time. Reviewer training designed and administered as a compliance checkbox (and lacking outcome measurement) are more dangerous than helpful, producing reviewers who satisfy an oversight presence requirement while avoiding substance.

Implementation guidelines. Effective reviewer training addresses three areas:
Decision support prepares reviewers to understand what the system is surfacing. Reviewers need structured presentations of the key factors in a decision, relevant case histories, and a basis for the system's recommendation. It should be noted that raw model outputs are rarely sufficient. Reviewer interfaces should present information in a format that supports judgment, extending beyond simple recording.

Signal recognition training prepares reviewers to interpret the specific signals the system generates: what a distributional novelty flag means in practice, how to interpret a low-confidence score, and how to identify when a case warrants escalation even if the system has not flagged it. Automation bias, the tendency to defer to algorithmic outputs even when independent judgment would differ, is a measurable phenomenon that targeted training can reduce.

Calibration closes the feedback loop. Reviewers should receive periodic data on the outcomes of decisions they reviewed: when they overrode the system recommendation, what happened? When they concurred, how did those cases resolve? Without outcome feedback, reviewers cannot improve their judgment, and the organization cannot distinguish reviewers who are adding value from those who are not.

Measurement indicator. Reviewers who have approved every case routed to them for six months without overrides or escalations is a red flag. In this situation, cases that do not require human judgment are subject to review (indicating the trigger threshold is misconfigured) or the reviewer is not applying independent judgment. Either condition should trigger a review of performance, training adequacy, or checkpoint calibration.

7.5 Measure Oversight Quality, Not Just Oversight Presence

High rubber-stamp rates with no overrides suggest reviewers are not adding value. Oversight presence (human review that exists in policy but not in practice) is inadequate and risky. The measurement framework should be designed to detect the difference and drive corrective action when the gap appears.

Implementation guidelines. Core metrics should be tracked at the decision-type level rather than the system level, since aggregate figures can mask significant variation across decision categories. The intervention rate measures what proportion of decisions of each type are being routed to human review. If a decision type that meets mandatory intervention criteria has a low intervention rate, this indicates the trigger is misconfigured. The override rate measures how frequently reviewers reach a different conclusion than the system recommendation. A sustained zero override rate is a signal that either reviewers are deferring uncritically, or the system is

performing exceptionally well, and the two conditions require different responses. Time-to-review tracks whether reviewers are completing reviews within the defined window and with enough time to exercise genuine judgment. Reviews completed in under 30 seconds for complex decisions are a warning indicator. Escalation frequency tracks whether reviewers are using the authority to escalate when cases warrant it.

Reporting requirements. In government settings, measurement outputs should be reported to program leadership on a defined schedule and included in annual AI governance reporting where required. When measurement data shows sustained rubber-stamp patterns, the corrective action should be documented: whether the response was reviewer retraining, threshold recalibration, or escalation of the concern to senior leadership. The measurement record itself is evidence of substantive oversight. An agency that can demonstrate it tracked oversight quality and acted on the results is in a materially stronger governance position than one that tracked only that reviews occurred.

7.6 Document Authority Structures and Accountability Records Explicitly

In government settings especially, the ability to demonstrate substantive human oversight to auditors, oversight bodies, and in legal proceedings depends on documentation that most current deployments do not maintain. Authority structure documentation and decision-level accountability records are the evidentiary foundation of AI governance.

Implementation guidelines. Authority structure documentation should identify, for each agentic deployment: the Human Oversight Lead by name and role; the reviewer roles authorized to review each decision type, including their qualifications and authority limits; the escalation chain above the primary reviewer level; and the change control process that governs modifications to any of these. This documentation should be treated as a governance artifact, reviewed annually, and updated whenever personnel or authority structures change.

Decision-level accountability records must go beyond logging that a review occurred. A complete accountability record identifies the reviewer by name, records their determination (e.g., concurrence, modification, or override) and documents the basis for the decision. In cases with legal significance a timestamp should establish when the review occurred relative to when the decision was executed. Anonymous review does not satisfy accountability requirements in consequential decision contexts. A checkbox that records only that a review was completed does not establish that the reviewer exercised judgment.

Government-specific requirement. Administrative and statutory frameworks in most jurisdictions require that decisions affecting individual rights and benefits be traceable to a named human decision maker who can be questioned, whose reasoning can be examined, and who can be held accountable for the outcome. An accountability record that names a reviewer but cannot establish what that reviewer knew, when they knew it, or what they determined provides limited protection in an audit or legal proceeding. Organizations building AI governance programs for the first time should work with agency counsel to identify the specific statutory and regulatory provisions that govern their decision types and ensure that accountability records satisfy those requirements.

The core premise of this paper is simple: human oversight of agentic AI is a structural requirement, not a procedural one. It must be designed in, not added on. The five decision characteristics identified here provide the operational basis for doing that design work rigorously and consistently across government and commercial settings. The six recommendations above translate that basis into practice.

About the Author and Initiative

Michael Bragen is Principal of ThinkCapital LLC and Director of the Government IT and AI Governance Initiative (GIAG), a practitioner research program examining the implementation of AI governance frameworks in government and commercial settings.

GIAG Stream Two focuses on human oversight quality in government agentic AI deployments, examining how agencies design, implement, and measure the human oversight functions embedded in AI-enabled workflows. Research findings are published at thinkcapital.org and in the Government AI in Practice newsletter.

Contact: michael.bragen@thinkcapital.org

Website: thinkcapital.org

Newsletter: thinkcapital.substack.com