TRANSFORMATION BOOK SERIES · No. 388

White Paper

BEYOND HUMAN APPROVAL

When the AI Agent Makes the Decision, Accountability Must Move Upstream

A decision-led governance model for agentic workflows in the first wave of enterprise deployment

Giovanni Leonardi

March 2024

When the AI Agent Makes the Decision, Accountability Must Move Upstream

Transformation · No. 388
Originally written March 2024 | Re-edited and re-rendered for the WhereBeyond Edition,

"Authority can be delegated to an agent; accountability cannot be delegated away."

1. The decision has already moved

At 09:12, an operations team authorises a language-model agent to reconcile a queue of supplier records. The design appears cautious: the agent may inspect documents, compare account data and recommend corrections. By lunchtime, however, the workflow has crossed a boundary nobody named. A confidence threshold automatically converts recommendations into changes. A second routine sends notices to suppliers. One disputed record has now become three organisational acts: a data amendment, a financial hold and an external communication.

Each technical component worked as configured. The failure was not hallucination in the narrow sense. It was that authority had accumulated across the workflow without a corresponding account of who owned the resulting decision.

This is the accountability gap emerging in the present wave of large-language-model experimentation. During 2023, many organisations treated agents as unusually capable assistants. In early 2024, tool use, retrieval and multi-step orchestration are turning the assistant into an actor. The important governance question is therefore no longer only, "Is the model output accurate?" It is, "What organisational decision has this system been allowed to complete?"

The evidence from early deployments points to one conclusion: governance must attach to the decision pathway, not merely to the model.

2. Why existing controls miss the problem

Most control environments were built for one of two familiar arrangements. Software executes deterministic rules, in which case testing and change control provide the main assurance. Or a person exercises judgement, in which case delegated authority,

supervision and professional accountability apply. An AI agent combines elements of both. It interprets ambiguous material, chooses a course, calls tools and may alter the environment in which the next choice is made.

That combination defeats three common approaches.

2.1 Model assurance is necessary but too narrow

Accuracy tests, bias reviews, prompt controls and security assessments matter. Yet a model can perform well on a benchmark while the end-to-end decision remains unsafe. The retrieval source may be stale. The tool may have broader permissions than the agent needs. An apparently reversible action may trigger a downstream process that is not reversible at all.

A model review asks whether the component behaves acceptably. Governance for agents must also ask whether the whole chain of action is legitimate.

2.2 Human approval can become ceremonial

A human-in-the-loop control sounds decisive until its operating conditions are examined. If one reviewer receives 480 agent recommendations during a two-hour window, approval is no longer meaningful judgement. It is throughput theatre. The person supplies a click, while the agent supplies the framing, evidence selection and recommended action.

The strongest argument for this arrangement is practical: human approval is familiar, auditable and easy to explain. That is true where the reviewer has time, competence, contrary evidence and genuine authority to refuse. Without those conditions, the control transfers the appearance of accountability to the reviewer without transferring the means to exercise it.

2.3 Policy statements do not define operating authority

Principles such as fairness, transparency and human oversight establish intent. They do not answer whether an agent may issue a refund, suspend an account, amend a forecast or contact a regulator. Teams then make local decisions in prompts, access settings and workflow code. Policy remains central while authority becomes distributed and largely invisible.

Authority can be delegated to an agent; accountability cannot be delegated away.

3. What the evidence is really showing

Across agent experiments, the recurring failures are less mysterious than they appear. They are usually created by the interaction of four conditions.

- The decision is unnamed. The initiative is described as document processing, customer support or workflow automation, so nobody states the judgement being delegated.
- Permissions are inherited. The agent receives the access of the service account or employee sponsoring the experiment rather than the minimum authority required for one task.
- Evidence is transient. Teams retain a final answer but not the retrieved material, tool calls, intermediate state and policy version that produced it.
- Escalation is defined by confidence alone. A numerical score becomes a proxy for consequence, even though a highly confident action can still have a serious impact.

Consider a composite procurement workflow tested in the first quarter of 2024. The agent reviews 1,200 low-value invoice exceptions each week. A pilot reports 92 per cent agreement with experienced analysts, which appears strong enough for automation. But the aggregate conceals the decision distribution. Eight hundred cases are clerical mismatches worth less than £100; the remaining four hundred include duplicate-payment warnings, disputed tax treatment and changes to supplier banking details. The error rate is concentrated in the smaller, consequential group.

A single accuracy figure therefore answers the wrong question. The useful evidence is segmented by decision class, consequence and reversibility. Once the team separates those dimensions, the control design changes: clerical corrections can proceed within a bounded threshold; bank-detail changes require independent verification; disputed tax cases cannot be completed by the agent at all.

The mechanism matters. Risk does not rise simply because an agent is more autonomous. It rises when uncertain interpretation is combined with consequential permission and weak recovery.

4. The accountability framework that is missing

A workable response begins with a decision register for agents. This is not an inventory of models. It is an inventory of organisational judgements and actions that agent-enabled workflows can make.

Decision class	Example	Permitted agency	Required evidence	Accountable role
Administrative	Classify and route a request	Execute within defined taxonomy	Source, classification and confidence	Process owner
Reversible commercial	Apply a low-value credit within policy	Execute within value and frequency limits	Policy version, calculation and tool log	Commercial owner
Material customer	Suspend access or reject a claim	Recommend only	Full case record and contrary evidence	Service decision owner
Regulated or fiduciary	Make a reportable determination	Prepare evidence; do not decide	Complete provenance and signed review	Named regulated function

The register creates a common object for business owners, risk specialists, architects and delivery teams. It also exposes where no accountable role exists. That absence must block deployment; it cannot be repaired by assigning the technology team after the event.

Four linked controls turn the register into an operating system.

4.1 Bound the authority

Each agent requires an explicit authority envelope: the decisions it may support, the actions it may complete, the values and volumes allowed, the systems it may touch and the duration of its mandate. Permissions should be issued for that envelope rather than inherited from a broad human role.

The envelope must also include prohibitions. “May propose a payment correction” is incomplete unless it also says “may not alter bank details, release funds or contact the supplier.”

4.2 Preserve the decision record

For every consequential action, retain enough evidence to reconstruct what happened. That normally includes the input, retrieved sources, relevant policy version, model and prompt configuration, intermediate tool calls, final action, human intervention and timestamp.

This is not a demand to preserve every hidden computation. It is a demand to preserve the organisational evidence that made the decision defensible. If the record cannot show which policy and source material governed the act, the organisation has an output log, not an audit trail.

© WhereBeyond, 2024–

4.3 Escalate by consequence and uncertainty

Confidence is one signal, not the governing rule. Escalation should combine uncertainty with materiality, novelty, affected party, cumulative volume and reversibility.

1. Low consequence, familiar and reversible: allow bounded execution with sampling.
2. Moderate consequence or unusual pattern: pause for informed review.
3. High consequence, irreversible or regulated: prohibit autonomous completion and route to a named decision owner.
4. Systemic anomaly: stop the workflow, preserve state and initiate incident handling.

This arrangement makes human involvement scarce and meaningful. It concentrates judgement where refusal, interpretation and accountability are real.

4.4 Design recovery before autonomy

An agent should not receive authority merely because it performs well in a test set. The organisation must first know how to detect drift, stop execution, reverse completed actions, notify affected parties and restore a trustworthy state.

A useful gate is simple: if the team cannot describe recovery for the maximum plausible action within the authority envelope, the envelope is too broad.

> *The unit of governance is not the prompt or the model. It is the decision, the authority attached to it and the evidence left behind.*

5. A recommendation for deployment decisions

Organisations should replace generic “human oversight” requirements with a decision-authority assessment at every agent deployment gate. The assessment should be owned jointly by the business decision owner and the technical service owner, with risk or legal participation where the consequence requires it.

Before an agent moves beyond experiment, the gate should require five artefacts:

- a decision register entry naming the judgement and action;
- an authority envelope with explicit limits and prohibitions;
- a segmented evaluation showing performance by decision class and consequence;
- a reconstructable evidence record;
- a tested stop-and-recovery procedure.

The responsible executive should approve the residual decision risk, not the novelty of the technology. Approval should expire when permissions, data sources, model configuration, policy or operating context changes materially. Agent governance is therefore a continuing mandate, not a one-time launch review.

There is a reasonable concern that this will slow useful automation. Poorly designed governance will. A committee reviewing every low-value action would reproduce the very bottleneck agents are meant to remove. The answer is proportionality: govern the decision classes once, automate within clear envelopes and reserve scarce human attention for exceptions with consequence.

That is faster than investigating an action nobody can reconstruct, reversing a chain nobody realised had begun, or asking a nominal reviewer to defend a decision they never truly made.

6. The test of accountable agency

The practical test is not whether an organisation can identify a person somewhere above the system. It is whether that person controls the conditions under which the agent acts: scope, evidence, escalation and recovery.

When those conditions are explicit, agentic workflows can remove clerical burden while preserving meaningful responsibility. When they are absent, apparent autonomy is simply dispersed authority. The model may be new, but the governance lesson is not. Decisions remain organisational acts, even when software assembles the evidence, chooses the tool and presses the button.

© WhereBeyond, 2024–