

BEYOND HUMAN SIGN-OFF

# When the Agent Acts: Governance for AI That Decides Rather Than Advises

Autonomous agents have crossed the line from recommending to acting — and the accountability model beneath enterprise transformation was never built for it.

Giovanni Leonardi

February 2024



# Contents

|                                     |   |
|-------------------------------------|---|
| 1. Executive Summary                | 3 |
| 2. The Overnight Reconciliation     | 4 |
| 3. Tracing the Roots                | 4 |
| 4. What Actually Changed            | 5 |
| 5. The Reflex, and Why It Half-Fits | 6 |
| 6. Toward a Settlement              | 7 |
| 7. The Deeper Pattern               | 8 |

# When the Agent Acts: Governance for AI That Decides Rather Than Advises

Transformation · No. 301

Originally written February 2024 | Re-edited and re-rendered for the WhereBeyond Edition,  
2025

---

*“Accountability, to mean anything, has to come to rest  
somewhere a summons can reach.”*

---

## 1. Executive Summary

---

Something quiet has changed in the way enterprises use software, and most governance functions have not yet noticed. For thirty years the machine recommended and a person decided; the flagged exception, the scored application, the suggested price all arrived at a human desk for a signature. The generative systems now being piloted under the banner of agents do not stop at the recommendation. Given a goal and a set of tools, they act — they raise the credit, clear the exception, send the message, book the refund. The decision has migrated from the person to the process, and the accountability model underneath enterprise change was built for a world where that migration had not happened.

This essay argues that the resulting gap is real but not unprecedented, and that seeing its precedents is the fastest route to closing it. We have delegated decisions to machines before — to credit scorecards, to program trading, to business-rules engines, to the robotic process automation of the last decade — and each time we struck a governance settlement to make the delegation safe. Those settlements rested on three quiet assumptions: that the machine's scope was narrow, that its behaviour was inspectable, and that a named human owned the rule it followed. The agent breaks all three at once. It is general rather than narrow, opaque rather than inspectable, and goal-seeking rather than rule-following, so there is often no rule and no owner to point to.

The reflex is to reach for what we already have — model risk management, the three lines of defence, a human in the loop — and each of those half-fits in a way that is more dangerous than an obvious misfit, because it looks like coverage. The honest position at the start of 2024 is that we do not yet have the settlement, only the outline of what it must do: govern the mandate rather than the model, keep the decision legible enough to reconstruct, calibrate control to reversibility rather than to the comfort of a signature, and restore a named owner to a decision the system now makes. None of that is a finished framework. It is the shape of the problem, traced back to its roots, so that the next few years of practice have somewhere sound to stand.

## 2. The Overnight Reconciliation

---

Consider a scene that would have been unremarkable eighteen months ago and is quietly extraordinary now. A finance team runs an intercompany reconciliation overnight. For years the routine was the same: a batch job matched what it could, and every unmatched item — a few hundred of them on a bad month — landed in a queue for an analyst to work through in the morning. The software's role ended at the boundary of the exception. It could sort, flag, and rank, but the judgement of what an unmatched item meant, and what to do about it, belonged to a person.

In the pilot now running, that boundary has moved. An agent built on a large language model is given access to the ledger, the email system, and the counterparty portal, and a plain-English instruction: clear the reconciliation. It reads the unmatched items, reasons about the likely cause, drafts and sends the query to the counterparty, interprets the reply, posts the correcting entry, and closes the item. On the first full run it cleared four hundred and ten of four hundred and sixty exceptions without a human touching them. The team's reaction was not relief but a particular kind of unease, and the unease is the subject of this essay. Someone asked the obvious question in the review meeting: if one of those postings is wrong, who made the decision to post it? Nobody in the room had a clean answer.

That question is not new. What is new is that we can no longer answer it with a name.

## 3. Tracing the Roots

---

It flatters the present to believe that delegating decisions to machines began with the language model. It did not. The interesting history of enterprise governance is largely a history of finding ways to let systems decide while keeping someone accountable for the deciding, and each wave left a settlement behind that is worth remembering precisely because we are about to need it.

The consumer credit scorecard is the cleanest early example. By the time it was mature, a statistical model was making millions of lending decisions that no human individually reviewed. This was decision-making by machine at industrial scale, and it was governed — not perfectly, but recognisably. The scorecard's scope was narrow and fixed: it decided one thing, accept or decline, within a defined product. Its logic, though statistical, was inspectable and stable; you could pull the model, examine the weights, and reproduce any decision it had made. And a named person — a chief credit officer, in the end — owned the policy the model expressed and answered for its outcomes to a regulator. The machine decided; the accountability stayed located.

Program trading told a similar story under more pressure. Algorithms placed orders in markets faster than any human could supervise in real time, and the governance answer was not to slow them to human speed but to bound them: position limits, pre-trade risk checks, kill switches, and circuit breakers that halted the machine when it left the envelope it had been given. Here the settlement was explicit about a truth the scorecard left implicit — you govern autonomous machine action not by watching each decision but by defining, in advance, the space inside which the machine is permitted to decide, and by being able to stop it.

Then came the business-rules engines and the straight-through processing of the 2000s, and the robotic process automation that swept back-office functions in the years since. RPA is the most instructive of all, because it is the most recent and because its governance is still fresh in the memory of everyone reading this. A software robot that reads an invoice and posts it to the ledger is making decisions a clerk used to make. We governed it, and we governed it by treating it as what it was: a deterministic script. Every branch it could take was written down. It did exactly what it was told, the same way every time, and when it failed it failed predictably and visibly, usually by stopping. The governance settlement for RPA was, in essence, the bot has no discretion, and because that was true, the accountability sat cleanly with whoever authored and owned the process.

*Across every prior wave, the settlement rested on the same three foundations: the machine's scope was narrow, its behaviour was inspectable, and a named human owned the rule it followed. Autonomous agents remove all three at once – which is why the old settlements half-fit, and why half-fitting is the dangerous case.*

Set those precedents side by side and the pattern is unmistakable. We have been governing machine decisions for a long time, and we have been reasonably good at it. The reason we were good at it is not that we watched every decision; at scale that was never possible. It is that the decisions were narrow, reproducible, and owned. Those were the load-bearing conditions. The agent kicks all three away.

## 4. What Actually Changed

It is worth being precise about the difference, because the temptation is to wave at the word autonomy and move on, and the precision is where the governance problem actually lives.

The first change is generality. A scorecard decided one thing; an RPA bot followed one script. An agent is given a goal and a box of tools, and it composes its own sequence of actions to reach the goal. You cannot enumerate its branches in advance because it writes them at runtime. The scope is no longer narrow, and worse, it is no longer fixed – the same agent handed a slightly different instruction will do something you never specified and never tested.

The second change is opacity. The scorecard was statistical but stable and inspectable; you could reproduce its decisions. A language model is neither transparent in its reasoning nor stable in its output. Ask it the same question twice and you may get two different answers, each plausible. The chain of reasoning it reports is a post-hoc narrative, not a reliable trace of how it actually arrived at the action. Inspectability, the second foundation, is gone.

The third change is the one that unsettled the finance team, and it is the collapse of the boundary between recommending and acting. Every prior system stopped at the edge of the decision or executed a decision a human had already encoded as a rule. The agent does neither. It decides and acts in one motion, at machine speed, and by the time a

human sees the result the action is already in the world. The refund is sent. The entry is posted. The message has left the building.

| Prior automation (scorecard, algo, RPA)     | Autonomous agent                     |
|---------------------------------------------|--------------------------------------|
| Narrow, fixed scope                         | General; composes its own actions    |
| Inspectable and reproducible                | Opaque and non-deterministic         |
| Follows a human-authored rule               | Pursues a goal; writes its own steps |
| Recommends, or executes an encoded decision | Decides and acts in one motion       |
| Named owner of the rule                     | No rule, and often no owner          |

Notice what this does to accountability. In every earlier settlement, if you asked who decided, you could trace a line to a person: the credit officer who owned the policy, the desk head who set the limits, the process owner who authored the script. The line was sometimes long, but it was continuous. With the agent, the line breaks. There is no rule whose author you can name, because the agent generated the behaviour. There is no reproducible decision to audit, because the decision was probabilistic and unlogged in any meaningful sense. The accountability has not moved to a new place. It has diffused, and a diffused accountability is functionally the same as none.

## 5. The Reflex, and Why It Half-Fits

Faced with this, the sensible institution reaches for the governance it already owns. This is the right instinct and the dangerous one, because the tools we own were built for the world the agent has just left, and each of them half-fits in a way that manufactures false comfort.

Take model risk management first, because it is the strongest candidate and the one most often proposed as the answer. Financial institutions have spent more than a decade building disciplined practices around models — validation, monitoring, documentation, independent review — under supervisory expectations that have been in force since the early 2010s. The argument runs: an agent is a model, we have a model risk framework, therefore we have a governance answer. Steelman it fully, because it is not foolish. Model risk management is genuinely mature, genuinely rigorous, and genuinely about machines that make consequential decisions.

And yet it was built for a specific kind of model: one with a defined input, a defined output, a measurable error, and a stable function you can validate once and monitor thereafter. Validation asks does this model predict well enough for its purpose? That question barely applies to an agent whose "purpose" is a natural-language goal and whose behaviour is a different sequence of tool calls every time. You cannot validate a decision boundary that does not exist. You cannot backtest an action space that is composed at runtime. Model risk management governs the model as an oracle that answers; it has nothing to say about the model as an actor that does. It fits the noun and misses the verb.

The three lines of defence has the same shape of failure. It is a sound structure, and it will remain part of the answer. But it presumes that the first line — the business — is where

decisions are made and owned, and that the second and third lines assure and audit that decision-making. When the decision is made by an agent, the first line's ownership becomes exactly the thing in question. You can assure a control you can locate; the difficulty is that the control has dissolved into the software the first line is supposed to own but did not write and cannot fully explain.

Then there is the most popular answer of all, the human in the loop. Keep a person in the decision, the reasoning goes, and accountability is restored: the agent proposes, the human disposes, the signature is back. If it held, it would dissolve the problem. It does not hold, for a reason we have known since long before language models, and the reason has a name in the older literature on automation: complacency. A human asked to approve decisions that are right ninety-eight per cent of the time does not stay vigilant for the two per cent. They become a rubber stamp, and everyone involved slowly comes to know it. The loop is present on the process diagram and absent in reality. Worse, the human-in-the-loop design often puts the person at precisely the point where they have the least ability to add value — reviewing an opaque recommendation at machine speed, with no realistic capacity to reconstruct how it was reached — while granting the comfort of believing a control exists. A control that is believed in but does not function is more dangerous than an acknowledged gap, because it stops the search for a real one.

---

*“We keep reaching for a human signature to close the gap, and the agent keeps making the signature ceremonial. The task is not to reinsert a person into the decision. It is to make the decision itself governable again.”*

---

## 6. Toward a Settlement

---

If the old tools only half-fit, the temptation is to demand a wholly new framework, delivered complete. That demand should be resisted, and not only because no one can honestly meet it at the start of 2024. It should be resisted because the precedents suggest the answer will be less a framework than a settlement — a set of practical accommodations that make the delegation safe enough to live with, arrived at by doing. What we can do now is name the properties that settlement will need, drawn from what actually made the earlier ones work.

The first is to govern the mandate rather than the model. Program trading points the way: we never governed the algorithm's internal reasoning, which we could not see; we governed the envelope it was allowed to act within. An agent should be given an explicit mandate — this agent may act on items under a stated value, in these systems, within these limits, and must escalate beyond them — and that mandate, not the model's weights, is the object of governance. The mandate is human-authored, inspectable, and ownable even when the model is none of those things. It restores the third foundation, ownership, by moving it from the rule the agent no longer follows to the boundary it must

not cross.

The second is to make the decision legible after the fact. We may have lost reproducibility, but we can insist on reconstructability: a durable, tamper-evident record of what the agent was asked, what tools it called, what it observed, and what it did, sufficient that a competent reviewer could later reconstruct the episode and judge it. This is not the same as the agent's self-reported reasoning, which is narrative. It is the flight-data recorder, not the pilot's account. Much of the governance of agents over the next few years will turn on the unglamorous discipline of logging, and institutions that treat this as an afterthought will find they have deployed decision-makers they cannot audit.

The third is to calibrate control to reversibility rather than to habit. Our instinct is to gate on the size of the decision — anything over a threshold gets a human. The more useful axis is whether the action can be undone. An agent that drafts a document a human will send has done something entirely reversible and needs little control; an agent that moves money, sends an external communication, or changes a customer's status has done something that cannot be recalled and needs a great deal. Sorting agent actions by reversibility, and reserving the scarce resource of genuine human attention for the irreversible ones, is a better use of governance than a uniform signature applied everywhere and meant nowhere.

- Mandate, not model — govern the explicit boundary the agent may act within, because it is human-authored and ownable even when the model is opaque.
- Reconstructable, not reproducible — keep a durable record sufficient to rebuild and judge any episode after the fact, distinct from the agent's own account of itself.
- Reversibility over threshold — spend real human scrutiny on the actions that cannot be undone, and let the reversible ones run.
- A named owner — every deployed agent has a person who answers for what it does, in the way a credit officer answered for the scorecard.

The fourth property is the oldest and the one most at risk of being lost in the excitement: a named owner. The scorecard worked as governance because a chief credit officer answered for it. The agent needs the same — not the vendor, not "the platform", not a diffuse committee, but a person whose responsibility it is that this agent, in this mandate, behaves. The technology has changed what the machine can do. It has not changed the fact that accountability, to mean anything, has to come to rest somewhere a summons can reach.

## 7. The Deeper Pattern

Step back far enough and the agent looks less like a rupture and more like the latest instance of a question the enterprise has been answering, and re-answering, for as long as it has used machines to decide: when the system acts, who is accountable? Each wave — the scorecard, the trading algorithm, the automated back office — forced the question again and produced an answer suited to its moment. The answers were never primarily technical. They were arrangements of scope, of visibility, and of ownership, wrapped around a technical capability to make it safe to use.

What unsettles us about the agent is that it arrived faster than the settlement, and it is being deployed into live processes while the governance is still a set of open questions. That gap between capability and control is not new either; it is the normal condition at the start of every automation wave, and it closes as practice accumulates. It closed for program trading after some painful lessons about what happens when the machine acts faster than the envelope can contain it. It will close for agents. The question is only how much is broken in the interval, and that depends almost entirely on whether the institutions deploying these systems treat governance as a brake to be released later or as part of the design from the first pilot.

The finance team in the opening scene did the right thing, though it did not feel like it. They noticed the unease. They asked who decided, found they could not answer, and treated the missing answer as a problem to solve rather than a detail to smooth over. That instinct — to insist that a decision the machine now makes must still trace to a person who owns it — is not nostalgia for a slower world. It is the thread that runs through every governance settlement we have ever made with a machine, and it is the thread most worth holding as the agents arrive. We are, once again, fluent in the technology well before we are fluent in the accountability. Closing that distance, deliberately and without waiting to be forced, is the transformation task that matters most in the year ahead.