

BEYOND THE AUTONOMY SLIDER

When to Let the Agent Decide

The trust calibration problem — and why the setting is a design choice, not a leap of faith

Giovanni Leonardi

November 2024



Contents

1. Executive Summary	3
2. The Year the Assistant Started Acting	3
3. Calibration, Not Capability	4
4. The Two Ways We Miscalibrate	5
4.1 Over-trust: the automation reflex	5
4.2 Under-trust: the supervision that eats the value	5
5. Why the Setting Drifts Wrong	6
6. A Grammar for Deciding What to Delegate	7
7. The Case for Never Letting Go — and Its Limits	8
8. What the Setting Reveals About Us	9

When to Let the Agent Decide

Transformation · No. 333

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

"A checkpoint that everyone has learned to click through is not a safeguard — it is a liability wearing the costume of one."

1. Executive Summary

In the space of roughly a year, the systems we deploy have crossed a quiet but decisive line. The question that occupied us until recently was whether a model could draft the message, reconcile the ledger, or triage the incoming request. That question is largely settled: it demonstrably can. The question now is whether we should let it send, post, or close — to act in the world without a person standing between the decision and its consequence. This is the trust calibration problem, and it is the defining transformation question of the moment.

The argument of this essay is that the difficulty organisations are having with agentic systems is not, at root, a technology problem. It is a calibration problem. The setting of how much autonomy to grant, on which decisions, under what conditions, is itself a design choice — and most organisations are making it by reflex rather than by design. They oscillate between two failure modes: an over-trust that hands the agent decisions it is not yet fit to own, and an under-trust that keeps a human on every step and quietly destroys the value the system was meant to create. Both are miscalibration. Both are expensive.

What follows examines why the setting so reliably drifts wrong, what structural forces sustain the drift, and how we might begin to treat delegation as a deliberate act of design rather than an act of faith. The uncomfortable conclusion is that where an organisation sets its trust dial reveals more about its temperament than about the technology in front of it — and that the gap between what these systems can do and what we actually let them do is where this year's transformation ambitions will mostly succeed or fail.

2. The Year the Assistant Started Acting

Picture the moment a capability crosses over. For most of the past two years the pattern was familiar and comfortable: a model suggested, a person disposed. The assistant drafted the reply and a human pressed send; it proposed the categorisation and a human confirmed it. The human was the load-bearing element, and everyone knew it. The system's confident errors were caught before they touched anything real, and so its

unreliability was, in a sense, free.

Then the frame shifted. The same underlying models were wired to tools — the ability to call an API, update a record, move money, operate a browser, execute a task end to end — and the assistant stopped merely advising and started acting. The demonstrations this year have been genuinely arresting: an agent that works through a multi-step task, decides what to do next based on what it finds, and completes it without being walked through each move. The word that attached itself to all of this, agentic, captured the shift precisely. The system now closes the loop itself.

And with that, the comfortable arrangement collapsed. The human who used to catch the confident error before it landed has been designed out of the path — that was the whole point, the source of the promised leverage. What we discovered, often in production rather than in the pilot, is that a system which is right most of the time but confidently wrong some of the time is a fundamentally different proposition once no one is checking. The error that used to be free now has a price, and the price is paid downstream, out of sight, at a moment no one chose.

The uncomfortable truth of agentic deployment is that removing the human from the loop does not remove the human's judgement from the system. It merely relocates that judgement to an earlier moment — the moment we decided how much to trust — and makes it invisible.

This is why so many agentic initiatives this year have stalled somewhere between an impressive demonstration and a dependable deployment. The demo shows what the system can do. The deployment forces the question the demo never had to answer: on which of these decisions are we actually willing to let it act alone?

3. Calibration, Not Capability

It is tempting to treat this as a capability gap — the models are not quite good enough yet, and when they improve the problem will dissolve. That framing is comforting and mostly wrong. Capability is improving quickly, but the calibration problem does not scale away with it; if anything, it sharpens. A more capable agent is trusted with more consequential decisions, so the stakes of each miscalibration rise in lockstep with the competence. We do not escape the problem by climbing the capability curve. We carry it with us.

The more useful reframing is to see trust not as a feeling we develop toward a system over time, but as a variable we set. Every agentic deployment embeds a threshold, whether or not anyone named it: this class of action, the system may take alone; that class, it may propose but not execute; this other class, it may not touch. The threshold exists the moment the system goes live. The only question is whether it was placed deliberately, with reasoning behind it, or inherited by accident from whatever the implementation happened to make easy.

“We are fluent in what the agent can do. We are far less fluent in deciding what it should be allowed to decide.”

That fluency gap is the real deficit. The engineering discipline required to build an agent that can act is now widely distributed. The organisational discipline required to calibrate what it acts upon is scarce, and it is a different kind of discipline entirely — less about capability and more about consequence, less about the model and more about the judgement we wrap around it. Calibration is the work of deciding, action class by action class, how the cost of a rare bad decision compares against the value of many good ones taken without friction. It is unglamorous, it produces no impressive demonstration, and it is where the value actually lives.

4. The Two Ways We Miscalibrate

Miscalibration is not a single error. It has two opposite faces, and an organisation can fail through either.

4.1 Over-trust: the automation reflex

The first failure is to grant autonomy the system has not earned on that particular class of decision. It usually arrives not through recklessness but through a kind of drift. The agent performs well through the pilot. Confidence accumulates. The exceptions that once triggered a human review are quietly folded into the automated path because reviewing them was slowing everything down and, after all, the system has been reliable. The threshold creeps upward, one reasonable-seeming decision at a time, until the agent is acting alone on precisely the consequential, ambiguous cases where its confident wrongness does the most damage.

This is not a new phenomenon; it is the automation complacency that industries dependent on autopilots and process control learned about decades ago, reappearing in a new medium. The mechanism is the same: reliable automation erodes the vigilance that was supposed to backstop it, and it erodes it most precisely when things are going well. What is new is the speed and the reach. A single mis-set threshold no longer produces one bad decision; it produces the same bad decision at machine scale, thousands of times, before anyone notices the pattern.

4.2 Under-trust: the supervision that eats the value

The opposite failure attracts far less attention because it looks responsible. Here the organisation, wary of the first failure, keeps a human confirmation on every action. Nothing the agent does executes without a person clicking approve. It feels prudent. It is often waste dressed as prudence.

Consider the arithmetic. If an agent handles a task in seconds but every output waits in a queue for a human to review and approve, the human review is now the constraint, and the system's speed advantage has been surrendered at the checkout. Worse, human review of high-volume, mostly-correct machine output degrades in a well-documented

way: presented with a long stream of outputs that are almost always fine, the reviewer's attention collapses into rubber-stamping. The oversight becomes ceremonial. The organisation is paying the full cost of human supervision and receiving almost none of its protective benefit — the worst of both settings at once.

Setting	What it feels like	What it actually costs
Over-trust	Efficient, modern, ambitious	Rare but severe errors, executed at scale, discovered late
Under-trust	Prudent, responsible, safe	The value surrendered to a queue; oversight that has quietly become a rubber stamp
Calibrated	Uneven, deliberate, occasionally awkward	The ongoing work of deciding, and defending, where each line sits

The point of setting the two failures side by side is that most governance conversations only guard against the first. They ask how do we stop the agent doing something terrible, and answer by adding human checkpoints — which is to say, they treat under-trust as free. It is not free. An organisation that reflexively supervises everything has not solved the calibration problem; it has chosen the other failure and congratulated itself on its caution.

5. Why the Setting Drifts Wrong

If calibration is so central, why do organisations so rarely do it well? The answer is that several structural forces all push in the same unhelpful direction, and understanding them matters more than any individual good intention.

The first is an accountability asymmetry. When an over-trusted agent causes harm, the failure is vivid, traceable, and owned by whoever authorised the autonomy. When an under-trusted agent quietly bleeds away its value through unnecessary supervision, no one is blamed, because nothing visibly breaks — the cost is an absence, a benefit never realised, and absences do not appear in incident reports. So the rational individual, protecting themselves, over-supervises. The organisation's aggregate caution is the sum of many people avoiding the blameable failure and ignoring the invisible one.

The second force is the demo-to-production chasm. The setting is almost always established during a period when the system is new, closely watched, and running on curated cases — exactly the conditions under which it performs best. Trust calibrated in that honeymoon does not survive contact with the long tail of real inputs, the adversarial edge cases, the malformed request, the deliberate attempt to manipulate the agent through its own inputs. The threshold was set against a distribution that production does not match.

The third is the hardest: most organisations cannot actually price the trade-off they are making, because it requires comparing quantities that live in different currencies. On one

side, the frequent, small, compounding value of decisions taken without friction. On the other, the rare, severe, hard-to-quantify cost of an autonomous error in a consequential case. There is no common unit. And in the absence of a common unit, the decision defaults to whichever consideration is culturally louder in that organisation – which is temperament, not analysis.

- In a cautious, control-oriented culture, the loud consideration is the catastrophic error, so the setting drifts toward paralysing supervision.
- In an aggressive, efficiency-oriented culture, the loud consideration is the productivity prize, so the setting drifts toward premature autonomy.
- In both, the drift is not a reasoned position. It is the organisation's disposition, expressed through a threshold no one deliberately placed.

There is one further force worth naming, because it is specific to this moment. The pace of capability improvement is itself destabilising the setting. A threshold that was correctly calibrated against last quarter's model may be wrong against this quarter's – in either direction – and few organisations have any mechanism for revisiting a delegation decision once it has been made. The setting is treated as a one-time configuration when it is in fact a standing judgement that decays.

6. A Grammar for Deciding What to Delegate

If the setting must be placed deliberately, we need a language for placing it. Not a rigid formula – the whole argument of this essay is against reflexive settings, and a formula is just a slower reflex – but a small set of questions that force the judgement into the open. Four dimensions do most of the work.

1. Reversibility. How easily can the action be undone if it was wrong? Reclassifying a document is nearly free to reverse; issuing a payment, publishing a public statement, or deleting a record is not. Reversibility is the single most useful axis, because it converts an unknowable question – will the agent be wrong? – into a manageable one: what does it cost us if it is, and can we recover?
2. Consequence. If this particular decision is wrong, how bad is the worst plausible outcome? An agent misrouting an internal query and an agent mishandling a safety-relevant or legally binding matter are not the same kind of decision, even if the model's confidence is identical in both.
3. Frequency and distribution. How often does this decision occur, and how varied are its cases? High-frequency, low-variance decisions are the natural home of autonomy; the value is large and the tail is thin. Low-frequency, high-variance decisions – the unusual, the ambiguous, the never-seen-before – are exactly where human judgement earns its cost.
4. Confidence legibility. Can the system tell us, in a way we can trust, when it is operating near the edge of its competence? An agent that reliably flags its own uncertainty can be granted autonomy on the confident majority while escalating the doubtful minority – which is a far more intelligent setting than a single global threshold applied to everything.

The move these four dimensions enable is the important one: to stop asking do we trust the agent? as though trust were a single dial for the whole system, and start asking do we

trust it with this class of action? as a separate decision for each class. A mature deployment does not have a trust level. It has a map — a deliberate assignment of autonomy that is generous where actions are reversible, frequent, and low-consequence, and conservative where they are irreversible, rare, and severe, with the system's own legible uncertainty routing the hard cases to people.

The right question is never "how much do we trust the agent" but "which decisions has it earned the right to make alone" — and that is not one question but as many questions as there are classes of decision.

None of this removes judgement from the loop. It relocates the judgement to where it belongs: to the deliberate, defensible act of drawing the lines, rather than to the accident of whatever the implementation made convenient.

7. The Case for Never Letting Go — and Its Limits

There is a serious position that says all of this is over-engineered, and that the safe answer is simply to keep a human confirmation on everything of consequence. Let the agent do the mechanical work; reserve every decision that matters for a person. It is the most defensible-sounding stance in the room, and it deserves a real answer rather than a dismissal.

Its strength is that it is honest about how bad the severe failures can be, and about how poor we are at anticipating them. If we genuinely cannot price the tail risk, the argument runs, the prudent move is to refuse the bet — keep the human, accept the lower ceiling on value, and sleep at night. In domains where the worst case is truly unbounded, this is not timidity; it is wisdom, and I would not want to talk anyone out of it there.

But as a general policy it fails, for two reasons. The first we have already seen: blanket supervision is not actually safe, because human review of high-volume machine output degrades into rubber-stamping. The stance promises protection it does not deliver; the human is present but no longer vigilant, so the organisation carries the full cost of oversight while its real protective value quietly erodes. A checkpoint that everyone has learned to click through is not a safeguard — it is a liability wearing the costume of one.

The second reason is competitive rather than technical. If a class of decision genuinely is reversible, frequent, and low-consequence, then refusing to delegate it is a standing tax — and in a market where others are calibrating rather than refusing, that tax compounds. The organisation that supervises everything does not merely forgo some efficiency; it slowly loses the ability to operate at the tempo the technology now permits. The blanket-supervision stance mistakes a real truth about the severe cases for a universal law, and in doing so it quietly chooses the invisible failure over the visible one. The honest position is not never let go. It is let go precisely where you can defend having done so, and nowhere else.

8. What the Setting Reveals About Us

Return, at the end, to where the trust dial actually gets set. Not in the model. Not in the tooling. In the temperament of the organisation doing the deploying — its appetite for the visible failure versus the invisible one, its capacity to hold two different kinds of cost in view at once, its willingness to make a judgement it will have to defend rather than hide behind a checkpoint or a blanket automation.

This is why the calibration problem is, in the end, a transformation problem and not a technical one. The pattern that recurs across the organisations struggling with agentic systems this year is not that their models are worse. It is that their institutional capacity to make and revisit a deliberate delegation decision is underdeveloped — and no amount of model improvement supplies it. The gap between what these systems can do and what a given organisation actually lets them do is a gap in organisational judgement, and it is precisely the gap between transformation intent and transformation reality that has defined every prior wave of technology adoption. The tools changed. The gap did not.

There is something clarifying in that. Agentic systems, by forcing us to state exactly which decisions we will and will not hand over, hold up a mirror. An organisation that cannot articulate why it drew a particular line has learned something uncomfortable about itself: that its caution, or its aggression, was never a considered position but a reflex it had mistaken for one. The work ahead is not primarily to build agents we can trust. It is to become the kind of organisation that knows what it is trusting them with, and can say why. The systems have started acting. Whether we have learned to decide what we are letting them decide is, still, the open question — and it is ours, not theirs, to answer.