

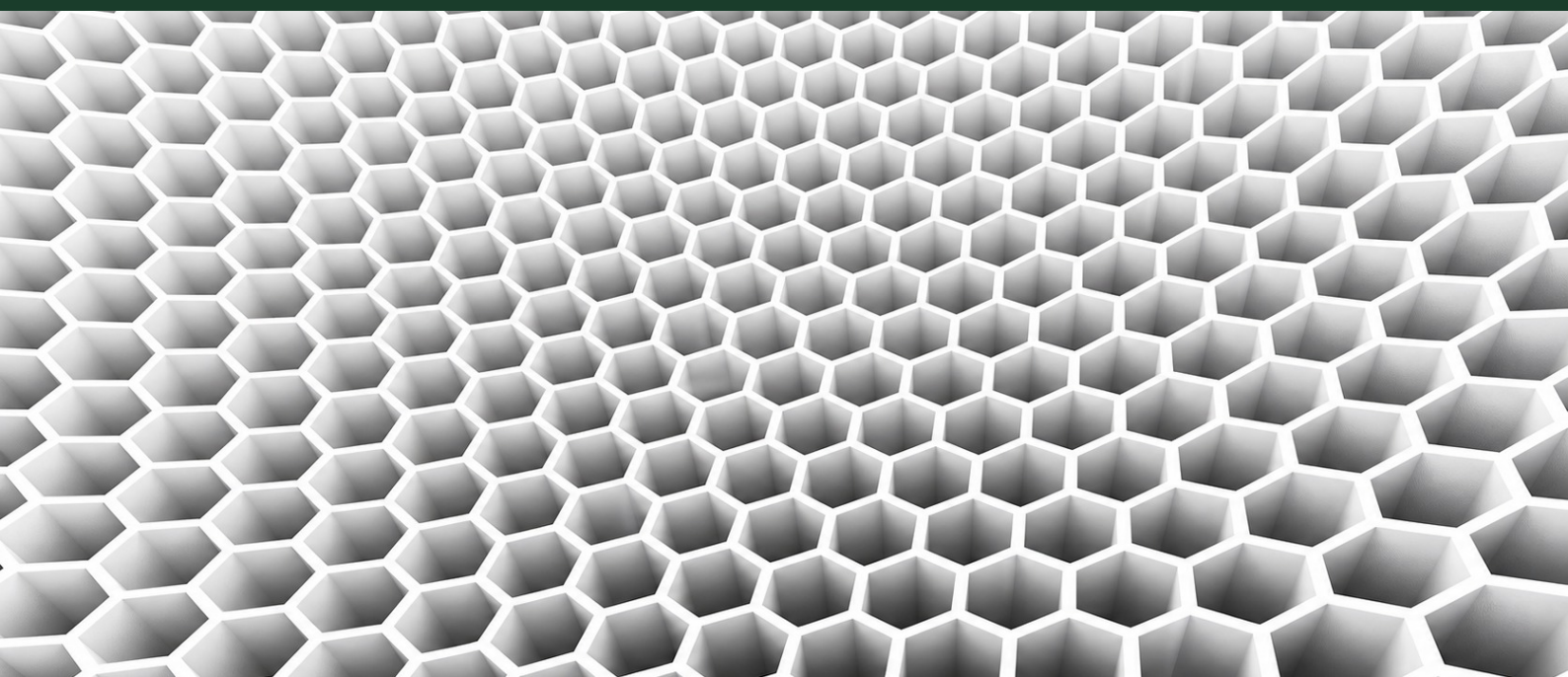
BEYOND CAPABILITY THEATRE

What Agents Cannot Do — A Field Inventory From Two Years of Practice

The Boundaries That Define the Human Role When the Mechanical Layer
Clears

Giovanni Leonardi

March 2026



What Agents Cannot Do — A Field Inventory From Two Years of Practice

Programme · No. 250
March 2026

“Wanting is not a cognitive function. It is the thing that keeps objectives alive when everything else conspires to let them quietly expire.”

1. The Demos and the Dread

The demo runs for forty seconds. An agent triages a portfolio of exception reports, cross-references three governance frameworks, and produces a stage-gate recommendation that took a team of four the better part of a week. The conference room applauds. Afterwards, in the corridor, the conversation turns to what it always turns to: how many roles this replaces.

I have spent the better part of two years deploying agent-based systems into programme and portfolio environments — orchestrating workflows, generating analysis, triaging exceptions, drafting governance artefacts. The question I keep returning to is neither what agents can do (which changes quarterly) nor what they might do (which is unknowable), but what they structurally cannot do. Not today's bugs. Not the limitations that will be patched in the next model release. The limits that persist because they are features of what agency means when it is artificial rather than human.

This inventory matters because it is not a list of weaknesses. It is a role specification — the most precise one available — for the human work that remains when the mechanical layer clears.

2. Transient Failures and Structural Boundaries

Every agent limitation I have encountered in practice falls into one of two categories. The first is transient: the agent misparses a date format, hallucinates a policy reference, fails to handle an edge case in a procurement workflow. These are real problems, sometimes expensive ones, but they have the character of engineering debt. They shrink. A team that catalogues them today will find half of them resolved in six months without any intervention on their part.

The second category is structural. These limitations do not shrink because they are not failures of capability but absences of something that capability cannot supply.

The organisations I work with tend to conflate the two categories. They treat every agent limitation as though it were a transient bug awaiting a fix, and build their human-role assumptions accordingly: the humans are here until the agents are good enough. This is precisely wrong. The structural limits are not waypoints on a road to full autonomy. They are the permanent geography of the terrain.

Five of them recur with sufficient consistency to constitute a field inventory.

3. Holding Accountability

An agent can execute a decision. It can apply a rule, weigh criteria, and select an option from a constrained set. What it cannot do is be answerable for the outcome.

This is not a philosophical abstraction. In programme governance, accountability is the mechanism by which organisations allocate consequence. When a senior responsible owner signs off a stage gate, they are not performing a cognitive task — they are accepting a burden. They are saying: if this goes wrong, the consequences flow to me. That acceptance changes behaviour upstream. It concentrates attention. It forces the question of whether the evidence is sufficient, not merely whether the evidence has been processed.

We discovered this concretely when an agent-generated stage-gate recommendation was adopted without meaningful human review. The recommendation was sound. The analysis was thorough. But when the programme encountered difficulty two months later, the governance review found no one who had owned the decision. The agent had done the thinking. No one had done the answering. The organisation had inadvertently created a gap in its accountability chain — not through negligence, but through a category error about what the agent's output represented.

Accountability requires an entity that can bear consequence. An agent can be retrained or decommissioned, but it cannot be held to account in any sense that creates the upstream behavioural effects that accountability is designed to produce.

4. Wanting Anything

An agent pursues objectives that are assigned to it. It does not hold them. The difference is operationally significant in any context where objectives must persist through ambiguity, resistance, and competing pressures.

In a complex transformation, the programme director who genuinely wants the outcome — who has staked professional reputation, career trajectory, and personal conviction on it — behaves differently from one who is merely executing a brief. They fight for resources in crowded portfolio meetings. They notice when energy drains from the room. They make the call to escalate before escalation is comfortable because they can feel the gap between where the programme is and where it needs to be.

An agent given the same objective will optimise diligently within its parameters. But parameters are not desire. When the context shifts, when the political wind changes, when a competing initiative quietly absorbs the oxygen, the agent does not notice that its objective is dying. It continues to optimise within its allocated envelope, reporting green while the ground shifts beneath it. We have seen this pattern repeatedly — agents maintaining governance rhythms and producing status assessments of increasing irrelevance because no one and nothing was holding the underlying intent.

Wanting is not a cognitive function. It is the thing that keeps objectives alive when everything else conspires to let them quietly expire.

5. Knowing What It Does Not Know

Calibration — the alignment between an agent's confidence and its actual accuracy — remains the least solved problem in deployed agent systems. And it fails in a specific direction: agents are confidently wrong rather than uncertainly right.

This is not a complaint about hallucination, which is a well-known and increasingly managed failure mode. It is a deeper problem about epistemic honesty at the boundary of competence. A seasoned programme manager, asked to assess a risk they have never encountered, will typically say so. They will flag that their judgement is extrapolated, that the analogies they are drawing are imperfect, that the confidence interval around their estimate is wide. This is not modesty. It is a professionally acquired skill — knowing the texture of one's own ignorance.

Agents do not have this skill, and the architecture of current systems makes it structurally difficult to acquire. The confident error — the precisely formatted, grammatically fluent, contextually plausible wrong answer — is the signature failure mode of agent-assisted governance. I have watched teams accept agent-generated risk assessments that were entirely invented, not because the teams were careless but because the output carried none of the signals that would normally indicate uncertainty. The absence of hedging was itself the deception.

Reliable self-knowledge about the boundaries of one's own competence is not a capability problem. It is a calibration problem, and calibration in the absence of genuine epistemic states — doubt, discomfort, the felt sense of being out of one's depth — may be structurally unachievable.

6. Reading the Room

In any organisation of sufficient complexity, the same correct answer can be wise in one context and catastrophic in another. The analysis that recommends consolidating two programmes is technically sound. Whether presenting it at this board meeting, to this sponsor, in this political moment, will advance or destroy the objective is a question that lives entirely outside the analysis.

Agents do not read rooms. They do not register that the finance director's silence means something different from the operations director's silence. They do not notice that the

steering committee's appetite for change has been exhausted by three consecutive difficult quarters, and that what the data supports and what the organisation can absorb are, this month, different things.

This is not emotional intelligence in the colloquial sense. It is political competence — the ability to map the informal power structures, the unspoken constraints, the history of relationships and betrayals that determine how a technically correct recommendation will actually land. Every experienced programme professional carries this map. It is rebuilt for each context, updated continuously through signals that are never written down, and deployed in real time to calibrate not what to say but when and how to say it.

Political context is not data. It is a continuously updated, embodied understanding of human dynamics that presupposes being a participant in those dynamics rather than an observer of them.

7. Bearing Novel Responsibility

The four limits above converge on a fifth, which is perhaps the most fundamental. Agents operate within the space of precedent — the patterns encoded in their training, the rules specified in their configuration, the boundaries set by their operators. When the situation is genuinely novel — when no precedent applies, when the rules conflict, when the right course of action must be invented rather than retrieved — the agent has nothing to draw on.

This is the domain of genuine judgement: the decision that cannot be justified by reference to what was done before, because nothing like this was done before. In programme environments, these moments are rarer than the literature suggests — most decisions do have precedent, and agents handle them competently. But the moments that matter most are disproportionately the ones that don't. The choice to kill a programme that the organisation has publicly committed to. The decision to restructure a governance framework mid-flight because the assumptions it was built on have been invalidated by events. The call to escalate past a sponsor who is part of the problem.

These decisions require someone who can stand in the space between the rules and act without them — and then live with having done so.

8. The Inventory as Role Specification

The argument against this inventory is that it is a snapshot — that structural limits are just transient limits on a longer timescale. Perhaps. But two years of practice have made me less persuaded by this objection, not more. The five boundaries I have described do not feel like engineering problems awaiting solutions. They feel like the outline of something that remains human precisely because it requires being human: being answerable, being driven, being honestly uncertain, being politically situated, being willing to act without precedent and bear the weight of having done so.

If that is right, then the inventory is the most practical tool the capability conversation has yet produced. It tells us what to select for — not residual expertise that agents have

not yet absorbed, but the capacities that define the human role regardless of what agents learn to do. It tells us what to develop — not AI literacy alone, though that matters, but the specific professional muscles of accountability, intent, calibration, political judgement, and novel decision-making. And it tells us what to protect — not jobs in the defensive sense, but the conditions under which these capacities can be exercised: governance structures that require human ownership, not merely human oversight.

“The capability conversation will continue to oscillate between demos and dread. The inventory offers a third register: a clear-eyed account of the boundaries, derived from practice, that tells organisations not what agents will eventually do but what humans must permanently be.”
