

BEYOND THE STAGE GATE

The Evaluation Bottleneck

A Practitioner's Perspective on the Shift from Production Scarcity to Evaluation Scarcity

Giovanni Leonardi

November 2025



The Evaluation Bottleneck

Portfolio · No. 118
November 2025

“The portfolio function was built for scarcity. Its next chapter must be written for abundance.”

1. The Inversion No One Planned For

For as long as portfolio management has existed as a discipline, its central problem has been scarcity. Scarce capital. Scarce talent. Scarce organisational attention. The portfolio office existed to allocate these scarce resources across competing demands, to ensure that the organisation invested in the right things, and to kill the initiatives that were not delivering. The entire apparatus — the stage-gate reviews, the benefits realisation frameworks, the quarterly prioritisation cycles — was built around the assumption that the constraint was capacity to do.

In 2025, that assumption is breaking.

The emergence of AI agents capable of generating code, analysis, design options, business cases, and strategic recommendations at unprecedented speed has introduced a problem that the portfolio management discipline was never designed to handle: the constraint is no longer capacity to produce, but capacity to evaluate. Organisations can now generate more options, more prototypes, more deliverables, and more analyses than their governance structures can meaningfully assess. The bottleneck has shifted from the supply side to the demand side — and the consequences are more significant than most portfolio leaders have yet recognised.

2. The Shape of the Problem

The pattern I have observed across multiple organisations in the past twelve months is remarkably consistent. A team deploys AI agents to accelerate a portion of their delivery pipeline — typically code generation, test automation, or document production. Productivity, measured in output volume, increases dramatically. The team celebrates. The programme board notes the improvement.

And then the problems begin.

- Review queues explode. Code review backlogs, which were already a source of friction in most engineering organisations, become unmanageable when AI agents can produce pull requests at ten times the previous rate. The human reviewers cannot keep pace. Quality assurance becomes the critical path — not because the work is bad, but because no one has the time to confirm whether it is good.

- Decision fatigue sets in at governance level. When an AI agent can produce five viable options for a strategic initiative in the time it used to take to produce one, the portfolio board must now evaluate five options instead of one. The cognitive load on decision-makers increases precisely when the pressure to decide quickly is also increasing. The result is either rushed evaluation or decision paralysis — both of which the portfolio function exists to prevent.
- Benefits realisation becomes opaque. If an AI agent has generated a business case, the portfolio office must now assess not just the merits of the case but the reliability of the analysis that produced it. Was the agent working from accurate data? Did it make assumptions that a human analyst would have challenged? The provenance of the analysis matters in ways it never did when a named human was accountable for the numbers.

3. Why the Existing Frameworks Are Insufficient

The standard portfolio management frameworks — and I include the major professional bodies' guidance here — assume a world in which the primary challenge is selecting the right investments from a manageable set of options, and then ensuring those investments deliver. The frameworks provide stage gates, scoring models, and prioritisation matrices. They are designed for a world of scarcity.

When the problem is not "we have too few options" but "we have too many outputs and insufficient capacity to evaluate them," the entire portfolio management apparatus needs to be rethought — not replaced, but fundamentally reoriented around evaluation capacity rather than investment selection.

This reorientation has several dimensions.

3.1 The Evaluation Bottleneck

The first and most immediate challenge is that evaluation is a human activity that does not scale the way production does. An AI agent can generate a hundred test cases in an hour. A senior engineer reviewing those test cases for correctness, relevance, and edge-case coverage cannot review a hundred in a day. The portfolio office must now treat evaluation capacity as a constrained resource and manage it accordingly — which means, paradoxically, that portfolio management needs more governance sophistication in the age of AI, not less.

This runs directly counter to the prevailing narrative, which holds that AI will simplify governance by automating reporting and analysis. It may automate the production of governance artefacts, but the judgement that those artefacts exist to inform cannot be delegated.

3.2 The Quality Assurance Paradox

The second dimension is subtler. When a human produces a deliverable, the organisation has some basis — however imperfect — for assessing confidence in the output. The

analyst's track record, the team's domain expertise, the review process that the work went through. When an AI agent produces the same deliverable, these proxies for confidence are absent. The output may be indistinguishable in form, but the organisation's ability to reason about its reliability is fundamentally different.

- The portfolio office is, in many organisations, the last line of defence against poor investment decisions. If it cannot distinguish between a rigorously produced analysis and a plausible-looking AI-generated one, its gatekeeping function is compromised.
- The temptation to treat AI-generated outputs as equivalent to human-generated ones — because they look the same — is exactly the trap that portfolio leaders must resist.

3.3 The Velocity Illusion

The third dimension concerns the relationship between speed and value. The pattern that recurs across complex programmes is this: an increase in output velocity creates the appearance of progress without necessarily creating actual progress. A programme that delivers twice as many features in a quarter has not necessarily delivered twice as much value. It may have delivered the same value with twice as much noise, or it may have delivered features that no one asked for because the agent was optimising for throughput rather than impact.

"The most dangerous outcome of AI-accelerated delivery is not that organisations build the wrong things. It is that they build so many things, so quickly, that they lose the ability to distinguish the valuable from the merely complete."

Portfolio management exists precisely to make this distinction. But its mechanisms were designed for a cadence that matched human production speed. When production speed increases by an order of magnitude, the portfolio function must either accelerate its evaluation capacity or accept that it is now operating with a significant lag — approving work that has already moved on, questioning decisions that have already been implemented, reviewing outputs that are already obsolete.

4. What a Portfolio Function Must Do Differently

The organisations that are navigating this well — and they are few, because the problem is new — share several characteristics.

1. They have separated evaluation from approval. The traditional stage gate combines evaluation ("is this good?") with approval ("should we continue?"). In a high-velocity environment, these must be decoupled. Continuous evaluation runs alongside production, with approval gates reserved for genuinely consequential decisions. This

is not a new idea — lean portfolio management has advocated for it — but AI-accelerated delivery makes it urgent.

2. They have invested in evaluation tooling. If AI agents are producing the work, then AI agents should also be producing the first-pass evaluation — flagging anomalies, checking consistency, identifying outputs that deviate from the brief. The human evaluator then reviews a curated set of concerns rather than the entire output. This does not eliminate the need for human judgement, but it makes human judgement tractable.
3. They have redefined portfolio metrics. Velocity is no longer a meaningful measure of portfolio health. The metrics that matter are evaluation throughput (how quickly can the organisation assess whether an output is valuable?), decision latency (how long between an output being produced and a decision being made about it?), and value density (what proportion of outputs are actually deployed, adopted, or acted upon?).
4. They treat AI-generated outputs with explicit provenance tracking. Every output carries metadata: which agent produced it, what data it was working from, what constraints were applied. This is not bureaucracy — it is the minimum viable information for a portfolio office to make an informed judgement about reliability.

5. The Larger Implication

The shift from production scarcity to evaluation scarcity is not a temporary adjustment. It is a structural change in how organisations create and govern value. Portfolio management, as a discipline, has an opportunity to become more important than it has ever been — because the need to distinguish signal from noise, value from volume, and progress from activity has never been greater.

But seizing that opportunity requires the discipline to move beyond its traditional comfort zone of investment selection and benefits tracking, and into the harder territory of evaluation design, quality assurance at scale, and the governance of AI-generated outputs.

The organisations that treat this as a technology problem — bolt on some AI review tools and call it done — will find themselves drowning in outputs they cannot assess. The ones that treat it as a governance problem — rethinking how decisions are made, how quality is assured, and how value is defined in a world of abundant production — will find themselves better positioned than they have been in years.

The portfolio function was built for scarcity. Its next chapter must be written for abundance.