

# Empiricism at Machine Speed

## ***Part 1 - The AI Governance Problem: Three Gaps That No Current Framework Closes***

---

*Matthew Hodgson, CEO, Zen Ex Machina & PST*  
*Dave West, CEO, Scrum.org*

**part 1 of 5**

## Executive Summary

Close to 80% of organizations have deployed generative AI in some form, yet only around 6% capture significant enterprise value from it (McKinsey & Company, 2025). The 5% that Boston Consulting Group classifies as future-built are not buying more AI than everyone else. They have redesigned how work happens around what AI can do (Boston Consulting Group [BCG], 2025). Because the leaders and the laggards have access to the same technology, what separates them is how the work around that technology is organized and governed.

Conventional project governance cannot supply that design. It was built for work with predictable effort, workers who flag their own uncertainty, and review that happens after a finished unit of work arrives. Agentic AI has none of those properties, so governance built on them leaves three questions without an owner:

- **Execution governance.** Who is accountable for what the AI produces?
- **Value governance.** Who decides whether it is working on the right problem?
- **Diagnostic governance.** What does the AI team's experience reveal about the organization's own structure?

The wider AI governance literature does not answer these questions either. Regulators write about compliance and liability. Technologists write about model safety and alignment. Both bodies of work matter, and both stop short of the day-to-day question of how a team that includes AI agents is run, with a person accountable for its output and its direction.

Scrum was designed for the same class of problem: complex work where the right answer cannot be specified up front. This series argues that Scrum addresses all three gaps, tests that argument against twelve months of operation in one agentic team, and names where the evidence stops. Papers 2, 3, and 4 provide the theory, the operational evidence, and the failure analysis.

## The Return Data Shows a Structural Problem, Not a Technology One

Most organizations now use generative AI, and few are getting enterprise value from it. Research aggregated across several datasets puts deployment at close to 80% of organizations and significant enterprise value at around 6% (McKinsey & Company, 2025). In McKinsey's analysis, broad use cases such as enterprise-wide copilots deliver diffuse gains that are hard to measure, while the higher-impact uses tied to specific functions mostly stall, with roughly 90% still sitting in pilot.

BCG's The Widening AI Value Gap surveyed 1,250 senior executives across 9 industries and assessed AI maturity on 41 dimensions. It found 3 performance tiers that are moving apart rather than converging (BCG, 2025). The models and platforms are available to every tier, so if the technology were the deciding factor, the tiers would converge as the technology spread. Because they are diverging instead, the difference has to lie in what each tier did to the organization around the tools.

**Table 1.** Three tiers of AI integration depth (adapted from BCG, 2025)

| <b>Share of organizations</b> | <b>Tier 1 - Reactive tool</b><br>~60%                                             | <b>Tier 2 - Productivity accelerator</b><br>~35%                                                                                                | <b>Tier 3 - Integrated Capability</b><br>~5%                                                                                  |
|-------------------------------|-----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| <b>What AI does</b>           | One step at a time, always started by a person. Full human control at every step. | Speeds up existing workflows. The underlying process design does not change.                                                                    | Agents run multi-step workflows on their own inside designed boundaries. Processes are redesigned around what AI can do.      |
| <b>Governance Model</b>       | None designed. Full human control substitutes for it.                             | Controls bolted onto the existing operating model: policy, access rules, review gates. The controls are designed. The model they sit in is not. | Governance and operating model designed together. Decision boundaries are revised each cycle based on what the cycle reveals. |
| <b>Return</b>                 | Individual efficiency gains. No enterprise-level advantage.                       | Capped at the existing process's ceiling.                                                                                                       | Returns grow with each operating cycle. Governance capability builds.                                                         |

On the same dataset, Tier 3 organizations generate 1.7 times the revenue growth of laggards, 3.6 times the 3-year total shareholder return, 2.7 times the return on invested capital, and 1.6 times the EBIT margin (BCG, 2025). They also turn pilots into production far more often: 62% of AI initiatives at future-built organizations reach production, compared with 12% at laggards. The speed at which a pilot becomes an operating capability is what separates the groups, and that speed depends on how the organization is structured rather than on what it spends.

The pattern predates any AI cycle. Brynjolfsson and McAfee (2014) found that general-purpose technologies do not deliver their economic potential until organizations restructure work around what the technology makes possible, and that using the technology to speed up existing tasks is not enough. Brynjolfsson, Rock, and Syverson (2021) described the shape of that transition as a productivity J-curve, in which performance dips before it rises because organizations must first pay for redesigned workflows and governance. The dip is the price of the redesign. Organizations that mistake it for a ceiling either retreat to tool-level use or push into governance models they have not yet built.

These figures establish a gap between adoption and value, and they locate it in how work is organized rather than in which tools are bought. They do not show which governance design closes it. Governance-specific data sharpens the question: 74% of companies plan to deploy agentic AI within 2 years, yet only 21% report a mature governance model for autonomous agents (Deloitte, 2026). Most organizations will therefore run autonomous agents under the governance they already have. The next question is whether that governance was built for this kind of work.

## **Conventional Governance Assumes Properties Agentic AI Does Not Have**

Project management governance rests on three assumptions: workers have steady cognitive capacity, effort predicts output, and workers flag their own uncertainty. Those assumptions hold well enough for people, which is why organizations can rely on staff to judge what they know and to escalate what they do not.

Agentic AI breaks all three. An agent's performance depends on how its context is structured, including where the relevant information sits within it (Liu et al., 2024), not on the hours it works. Output quality

depends on how the task is framed: meaning-preserving changes to prompt formatting alone can shift a model’s accuracy substantially (Sclar et al., 2024). And agentic systems do not signal uncertainty in any way that can stand in for human judgment; methods for eliciting a model’s confidence remain unreliable, particularly on tasks that require professional knowledge (Xiong et al., 2024). Speed makes all three worse: agents can produce work far faster than people can review it, and because queue length drives lead time (Reinertsen, 2009), sequential approval gates turn from a safeguard into a bottleneck.

Organizations facing this mismatch tend to respond in a predictable way. Snowden and Boone (2007) call the response *entrained thinking*: applying methods that worked in familiar conditions to situations that call for different ones. Faced with agentic AI, organizations reach for what has always worked: stakeholder consultation, requirements definition, role definition, and approval gates. That conditioned behavior reduces the discomfort of not knowing, at the cost of adapting to what is actually in front of the organization. The result is governance designed for certainty, applied to work that offers none.

Project governance leaves three questions without an owner, shown in the last three rows of Table 2: who holds accountability for direction, how execution stays visible without per-task oversight, and what happens to impediments.

**Table 2.** Where the assumptions break

| Dimension             | Project management governance                                    | What agentic AI requires                                                                             |
|-----------------------|------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| Worker type           | People with steady cognitive capacity                            | Agents whose capacity varies with context and instruction structure, not hours                       |
| Effort and output     | Output predicted from task size and experience                   | Not predictable from task size; quality is sensitive to how the task is framed                       |
| Signaling uncertainty | Workers ask, clarify, or escalate                                | Agents produce no confidence signal that replaces human judgment                                     |
| Quality assurance     | Review after a finished unit of work (ISO 9001, CMMI, Six Sigma) | Production outruns review; sequential gates create the bottleneck the automation was meant to remove |
| Response to novelty   | Define the workflow, the roles, and the approvals                | Complex work needs inspect-and-adapt governance, not certainty-domain logic                          |
| Value accountability  | Sponsor accountable for budget; value fixed at inception         | A single accountable owner for direction, which cannot be held by a committee                        |

| Dimension            | Project management governance                                              | What agentic AI requires                                                           |
|----------------------|----------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| Execution visibility | Milestones, status reports, task completion                                | An Increment that is visible and inspectable without per-task oversight            |
| Diagnostic function  | Impediments treated as production problems and absorbed as normal friction | Impediments arrive as named, dated, traceable evidence of what the structure costs |

Sources: Liu et al. (2024); Reinertsen (2009); Schwaber & Sutherland (2020); Sclar et al. (2024); Snowden & Boone (2007); Xiong et al. (2024).

## Three Gaps Open Where the Assumptions Break

Each broken assumption leaves a question that nobody in a conventional governance structure is set up to answer, and the wider AI governance literature does not answer them either. Regulators address compliance frameworks, liability, and the duties of organizations that deploy AI at scale (European Parliament and Council of the European Union, 2024; OECD, 2024). Technologists address model safety, alignment, and the controls that keep an agent's operations safe and accountable (Shavit et al., 2023). Both bodies of work matter, and neither reaches the operational question a practitioner faces every day.

That question is not "How do we stop the AI from doing something catastrophic?" It is "How do we run a team that includes AI agents, reliably, at quality, with a person accountable?" Compliance frameworks and alignment research do not answer it, because the answer lies in how the team is governed. The three gaps below are the parts of that question left open.

### Gap 1 — Execution governance

Execution governance follows from the broken assumptions about effort and uncertainty. When effort no longer predicts output and the worker cannot flag its own doubt, quality can no longer be inferred from who did the work or how long it took. Someone therefore has to be accountable for how an AI agent operates within a production workflow, how quality is ensured, and how errors are caught, and every output has to be traceable back to the instruction that produced it.

An output that meets the required standard, but whose production logic cannot be reconstructed, does not satisfy accountability, because an organization that cannot say which instruction produced a result cannot defend it. Governance that cannot be audited fails the test it exists to meet.

Shavit et al. (2023) propose that at least one human party be accountable for any harm an agent causes, and that an agent's actions be attributable and recorded. That is accountability for harm. It does not tell a team who is accountable for whether each output meets a quality standard. The exposure is broad: with only 21% of organizations reporting a mature governance model for autonomous agents (Deloitte, 2026), most of the agentic deployments planned for the next 2 years will begin without one.

## **Gap 2 — Value governance**

Value governance follows from how project governance treats direction. A sponsor is accountable for budget, and value is fixed at inception (Table 2). Execution governance asks whether the AI is working correctly. Value governance asks whether it is working on the right thing, and that second question comes first.

An AI system that executes flawlessly on the wrong problem is a governance failure. The choice of problem matters more than the quality of the solution (Gothelf & Seiden, 2021). A perfect answer to the wrong question costs money twice: once to build it, and again for the work the team did not do instead.

Without an explicit owner, direction defaults to whoever configured the system last, and current AI governance guidance leaves that default in place. The OECD (2024) principles and the EU AI Act (European Parliament and Council of the European Union, 2024) address risk, transparency, human oversight, and accountability for compliance. Neither assigns anyone accountability for direction, meaning the decision about which problem an AI system should be working on.

## **Gap 3 — Diagnostic governance**

Diagnostic governance is the least visible of the three. It follows from how project governance treats impediments: as production problems, absorbed as normal friction (Table 2). Most frameworks also treat transparency as one-directional, with information flowing up from the AI system to the humans above it. That is correct and incomplete.

Scrum's empiricism rests on transparency, inspection, and adaptation (Schwaber & Sutherland, 2020). In a system built on those pillars, transparency is a sensing mechanism as much as a reporting one. What an AI team cannot do reveals what the organization has decided, by action or by inaction, not to change. Without a mechanism that records those limits, each one is absorbed as friction, and the organization loses the evidence of what its own structure costs.

## The Gap Widens as AI Maturity Rises

The three gaps matter little while a person stands at every step, and they become decisive once that person steps back. At Tier 1, full human control substitutes for governance design, because a person starts and checks every interaction. At Tier 2, controls are bolted onto the existing operating model and each output is still reviewed individually, so the gaps stay covered, at the cost of capping returns at the existing process's ceiling. At Tier 3, agents perform multi-step work on their own and no person is present at each step. Governance then has to be designed, because nothing else holds the system together.

The most substantial practitioner guidance for that setting is Shavit et al.'s (2023) OpenAI paper, which sets out practices for keeping agents' operations safe and accountable, from evaluating whether an agent is suitable for a task to keeping its actions legible and interruptible. Those practices assume that someone can judge whether the agent's output is correct. In software development, automated tests often supply that judgment.

Knowledge work cannot be verified that way. An executive briefing, a research synthesis, a piece of strategic advice, or a policy analysis is judged by its audience and purpose. No test suite returns pass or fail. That is the setting this series addresses: governance for agentic knowledge work, where quality is judged rather than tested.

## Scrum Addresses All Three Gaps

Scrum was designed to generate adaptive solutions to complex problems (Schwaber & Sutherland, 2020): problems where cause and effect are only clear afterward (Snowden & Boone, 2007), and where the right answer cannot be specified in advance. Its design is the structural opposite of the certainty assumption that makes project governance fail with agentic AI, and each gap maps to a mechanism Scrum already has.

For execution governance, the Definition of Done sets the quality standard every Increment must meet, and the Developers are accountable for adhering to it (Schwaber & Sutherland, 2020). Sprint events make the work visible and inspectable at a regular cadence, so quality is governed without per-task oversight. Accountability is assigned before the work starts rather than reconstructed after something goes wrong.

For value governance, a Product Owner commits to a Product Goal. That commitment turns an investment decision into an explicit standard that every later decision answers to. Without it, value stays a preference and is settled by whoever configured the system last. With

it, value becomes a constraint the organization can trace, revise, and inspect at every Sprint Review. The Product Owner is one person, not a committee (Schwaber & Sutherland, 2020), so the accountability for direction that current AI guidance leaves unassigned has a named owner.

For diagnostic governance, the empirical pillars turn impediments into evidence. The evidence in this series comes from one agentic team operating inside Scrum for twelve months, reported in Paper 3. Consider an impediment of the kind that evidence examines: a batch approval queue that blocks three Increments across a single Sprint. That is a throughput finding with a cost attached and a structural cause, and it arrives named, dated, and traceable to a decision someone made. Informal escalation produces nothing like that, and governance bodies find an informal complaint far easier to dismiss.

Scrum does not guarantee that leadership acts on what a Sprint Review surfaces. It guarantees that the evidence is produced on a regular cadence, in a form that can be inspected and attributed, and added to over time. Table 3 summarizes the mapping.

**Table 3.** The three gaps and what addresses them

| Gap                   | Why it stays unaddressed                                                                                                                                    | What Scrum supplies                                                                                                                                                                |
|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Execution governance  | Current guidance assigns accountability for harm, not for whether each output meets a quality standard or can be traced to the instruction that produced it | The Definition of Done sets the quality standard, the Developers are accountable for meeting it, and Sprint events make AI work visible and inspectable without per-task oversight |
| Value governance      | Direction defaults to whoever configured the system last                                                                                                    | The Product Owner and Product Goal turn direction into an explicit commitment that can be audited                                                                                  |
| Diagnostic governance | Impediments get absorbed as normal friction, with nothing to convert them into evidence                                                                     | The empirical pillars turn Sprint impediments into a dated record of what the structure costs                                                                                      |

Paper 2 sets out why the mapping holds. The practical consequence for this paper is that organizations do not need to gather requirements, invent roles, or write a new process for agentic AI, which is the entrained response described earlier. They need a framework built for complex adaptive work, and organizations that already run Scrum in their delivery teams have one.

it, value becomes a constraint the organization can trace, revise, and inspect at every Sprint Review. The Product Owner is one person, not a committee (Schwaber & Sutherland, 2020), so the accountability for direction that current AI guidance leaves unassigned has a named owner.

For diagnostic governance, the empirical pillars turn impediments into evidence. The evidence in this series comes from one agentic team operating inside Scrum for twelve months, reported in Paper 3. Consider an impediment of the kind that evidence examines: a batch approval queue that blocks three Increments across a single Sprint. That is a throughput finding with a cost attached and a structural cause, and it arrives named, dated, and traceable to a decision someone made. Informal escalation produces nothing like that, and governance bodies find an informal complaint far easier to dismiss.

Scrum does not guarantee that leadership acts on what a Sprint Review surfaces. It guarantees that the evidence is produced on a regular cadence, in a form that can be inspected and attributed, and added to over time. Table 3 summarizes the mapping.

**Table 3.** The three gaps and what addresses them

| Gap                   | Why it stays unaddressed                                                                                                                                    | What Scrum supplies                                                                                                                                                                |
|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Execution governance  | Current guidance assigns accountability for harm, not for whether each output meets a quality standard or can be traced to the instruction that produced it | The Definition of Done sets the quality standard, the Developers are accountable for meeting it, and Sprint events make AI work visible and inspectable without per-task oversight |
| Value governance      | Direction defaults to whoever configured the system last                                                                                                    | The Product Owner and Product Goal turn direction into an explicit commitment that can be audited                                                                                  |
| Diagnostic governance | Impediments get absorbed as normal friction, with nothing to convert them into evidence                                                                     | The empirical pillars turn Sprint impediments into a dated record of what the structure costs                                                                                      |

Paper 2 sets out why the mapping holds. The practical consequence for this paper is that organizations do not need to gather requirements, invent roles, or write a new process for agentic AI, which is the entrained response described earlier. They need a framework built for complex adaptive work, and organizations that already run Scrum in their delivery teams have one.

## What This Means for Leadership

There are three implications for an executive responsible for AI investment:

**Accountability for AI output is an organizational design decision.** Execution governance needs accountability that is explicit and assigned before the work begins: a team accountable for meeting a Definition of Done, and a single Product Owner accountable for the value of what the team produces. A committee, working group, or steering board cannot hold either accountability. Where the current operating model has no one in those positions, the gap is a design decision for the organization rather than a technology setting to adjust.

**Value governance happens earlier than most AI conversations reach.** “Is the AI working on the right problem?” is first answered before the first Sprint begins, and it is revisited at every Sprint Review after that. A Product Goal turns the default into an explicit commitment that can be revised and inspected. Without one, inertia sets the direction.

**The diagnostic function is the most easily overlooked and the most structurally valuable.** An AI team inside an empirical framework produces systematic evidence about what an organization’s structure and incentives have quietly decided not to change. That evidence accumulates across Sprints, dated and traceable to specific decisions. What an organization does with it remains a leadership question, but the organization can no longer say it did not know.

Organizations moving toward Tier 3 will meet all three gaps whether or not they have designed for them. The choice is whether they meet them with governance built for the purpose.

## Conclusion

Organizations deploying agentic AI are not failing because their people lack skill or their technology lacks sophistication. The return data locates the problem in how work is organized, and this paper locates it more precisely: governance built for ordered, predictable work has been pointed at work that is neither. Adding process within that governance model leaves the assumptions it rests on untouched, so the gaps stay open.

Scrum is designed to address all three gaps because it was built for this class of problem. The Product Owner and Product Goal make direction explicit and accountable. The Definition of Done and Sprint events make autonomous work visible and its quality governable without per-task oversight. The empirical pillars turn the impediments that frustrate most deployments into a dated record of what the organization’s structure actually costs.

The framework already exists. What the evidence shows so far is that it held in one agentic team over twelve months. Whether it holds in regulated settings and across many teams is the test ahead. The decision is whether to apply it and learn, or to keep adapting frameworks built for a different problem and expect a different result.

## About the Authors

**Matthew Hodgson** is co-founder and CEO of Zen Ex Machina, a diagnostic-first executive strategy and advisory firm operating across government and corporate Australia and the Asia-Pacific region. He is a Scrum.org Professional Scrum Trainer and has led complex organizational transformations across Australian federal agencies and corporate environments for over 20 years. He is the author of *Evolve: The Operating Model AI Demands*.

**Dave West** is CEO at Scrum.org, The Home of Scrum. A frequent keynote speaker and widely published author, he co-authored *The Nexus Framework for Scaling Scrum* and *Head First Object-Oriented Analysis and Design*. He led development of the Rational Unified Process at IBM/Rational, managed Ivar Jacobson Consulting for North America, and served as VP research director at Forrester Research. Before Scrum.org, he was Chief Product Officer at Tasktop, helping grow it into a VC-backed company of nearly 100.

## About Scrum.org

Scrum.org, the Home of Scrum, was founded by Scrum co-creator Ken Schwaber as a mission-based organization to help people and teams solve complex problems. We do this by enabling people to apply Professional Scrum through training courses, certifications and ongoing learning all based on a common competency model.

## References

Boston Consulting Group. (2025). *The widening AI value gap: Build for the future 2025*. Boston Consulting Group.  
<https://www.bcg.com/publications/2025/are-you-generating-value-from-ai-the-widening-gap>

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton & Company.

Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1), 333–372.  
<https://doi.org/10.1257/mac.20180386>

Deloitte. (2026). *State of AI in the enterprise: The untapped edge*. Deloitte AI Institute. <https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html>

## References Continued

European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>

Gothelf, J., & Seiden, J. (2021). *Lean UX: Designing great products with agile teams* (3rd ed.). O'Reilly Media.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173. [https://doi.org/10.1162/tacl\\_a\\_00638](https://doi.org/10.1162/tacl_a_00638)

McKinsey & Company. (2025, March 12). *The state of AI: How organizations are rewiring to capture value*. McKinsey & Company. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>

OECD. (2024). *OECD AI principles overview*. OECD.AI Policy Observatory. <https://oecd.ai/en/ai-principles>

Reinertsen, D. G. (2009). *The principles of product development flow: Second generation lean product development*. Celeritas Publishing.

Schwaber, K., & Sutherland, J. (2020). *The Scrum guide: The definitive guide to Scrum — The rules of the game*. Scrum.org. <https://scrumguides.org>

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. *In Proceedings of the 12th International Conference on Learning Representations*. <https://arxiv.org/abs/2310.11324>

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Beutel, A., Passos, A., & Robinson, D. G. (2023). *Practices for governing agentic AI systems*. OpenAI. <https://openai.com/index/practices-for-governing-agentic-ai-systems/>

Snowden, D. J., & Boone, M. E. (2007). *A leader's framework for decision making*. Harvard Business Review, 85(11), 68–76.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. *In Proceedings of the 12th International Conference on Learning Representations*. <https://arxiv.org/abs/2306.13063>