Agentic Governance Will Not Come from the Models

Action-space, autonomy, reversibility and decision regimes: toward an exogenous governance architecture for agentic AI systems

Jérôme Vetillard / Twingital Institute® - April 2026

Abstract

Agentic AI systems reached, in 2026, a threshold of industrial deployment that makes the question of their governance urgent. This article argues that the governability of an agentic system cannot be expected as an emergent property of model progress alone. It must be produced by the architecture of the system that puts them in a position to act. The article proposes a risk grid structured around three axes: the extent of the action-space, the degree of autonomy, and the reversibility of actions, supplemented by two modulators, domain criticality and error asymmetry. This grid grounds a taxonomy of three decision regimes: assistance, structured recommendation, and bounded execution, which do not form a maturity hierarchy but distinct architectural modes. The article then examines the pattern of decision contracts, derived from design-by-contract logic, as an operational admission control mechanism. The emerging institutional frameworks from Singapore, the United States, the European Union and the United Kingdom are analyzed. Six limitations and a four-question research program are explicitly formulated.

1. The Problem: More Capable Agents, Not Governable by Construction

The central problem of agentic AI is not that agents are becoming more capable. It is that they are becoming actionable before they are governable.

Agentic AI systems reached, in 2026, a threshold of industrial visibility that demands a specific architectural examination. Market analysts anticipate a rapid diffusion of task-specific agents within enterprise applications over the coming years. Gartner projects that up to 40% of enterprise applications will integrate task-specific agents by the end of 2026, up from less than 5% in 2025. This estimate belongs to market forecasting rather than consolidated empirical observation, but it signals a clear strategic direction: the agent is no longer merely a technical demonstration. It is becoming a plausible, and progressively commonplace, component of contemporary information systems.[1]

This surge in capability, however, resolves nothing of the principal challenge. An agentic system is not merely a system that produces text, code or a classification. It is a system that can, in certain configurations, select tools, mobilize data, trigger actions, chain operations, revise its plan and, sometimes, act upon artifacts or environments external to the model itself. The problem therefore does not reside solely in the quality of the output. It resides in the decision regime introduced by the architecture.

The shift is decisive. In a deterministic workflow, the control graph is defined in advance: a component executes at a predetermined point, in a predetermined order, with a relatively stable scope of effect. In an agentic system in the strong sense, the component has partial latitude over the choice of steps, tools or action sequences. The system no longer merely executes a path; it explores a decision space. This is not a terminological nuance. It is a regime change. And it is precisely what makes agentic governance distinct from classical software governance.

This distinction is all the more important given that recent gains in capability should not be conflated with equivalent gains in reliability. The work of Rabanser, Kapoor, Kirgis, Liu, Utpala and Narayanan proposes a structured evaluation framework around four dimensions (consistency, robustness, predictability and safety) based on twelve concrete metrics. Their contribution is not to demonstrate a theoretical impossibility of making agents reliable, but to document an empirical gap: observed capability gains do not mechanically translate into substantial and homogeneous gains across the reliability dimensions most pertinent in operational contexts. In other words, the fact that an agent can do more does not yet mean that it does what it does in a stable, reconstructible and predictable manner.[2]

The present article argues, from this observation, a simple thesis. The governability of an agentic system cannot be expected as an emergent property of model progress alone. It must be produced by the architecture of the system that puts them in a position to act. This governance will be called here exogenous, not because it would ignore the internal properties of the model, but because it does not depend on them to exist. By exogenous governance, we designate the set of control, validation, traceability, bounding and audit mechanisms that are external to the model itself: declarative policies, admission controllers, decision contracts, human checkpoints, separation of action rights, allowlists, evidence logs and revocation mechanisms. The property of governability is then not a psychological quality attributed to the model; it becomes a property of the system.

One must, however, avoid artificial dramatization. Not everything that presents itself today as an "agent" is an agent in the strong sense. Simon Willison describes a landscape where genuinely useful applications often rely on framed and heavily instrumented engineering patterns, notably in code agents, rather than on open and undifferentiated autonomy. This observation remains that of a practitioner, not a market survey. But it carries useful analytical reach. It reminds us that the problem addressed here concerns less the totality of enterprise LLM uses than the fraction of systems in which a real decision space is granted to the component. It is precisely this fraction, still a minority by volume but structurally significant in terms of risk, that most strongly concentrates the governance stakes.[3]

The question is therefore not whether models are becoming more impressive. It is whether a system that delegates a share of initiative to them can remain opposable, explainable, revisable and accountable when its decisions produce effects on third parties. It is at this level that governance ceases to be a compliance supplement and becomes an architectural problem.

2. The Counter-Thesis: Will Endogenous Model Improvement Render Architectural Safeguards Marginal?

The thesis of strong exogenous governance must take its best objection seriously. This objection can be formulated as follows: models progress from generation to generation, alignment techniques reduce certain forms of behavioral deviation, outputs become better structured, benchmarking tools multiply, and one can envisage that as measured reliability increases, the level of autonomy conceded to the system might be dynamically adjusted. In this perspective, the heavy architectural governance mechanisms would be merely a transitional state. The system would become progressively governable because its central component would itself become more stable, better calibrated, better aligned and better monitored.

This position should not be caricatured. It rests on a real empirical trajectory. Recent models are, in many respects, more useful and more structured than their predecessors: system cards and evaluations published by several laboratories document measurable gains in instruction-following, output structuring and constrained format management.[10] It would be absurd to deny the value of these endogenous advances. The question is therefore not to dogmatically oppose them to all architectural governance. The question is whether they suffice.

Three reasons lead to a negative answer.

The first concerns the distinction between capability and consistency. Capability can grow with scale, data quality, training sophistication or alignment methods. Consistency, in the sense of behavioral stability on similar tasks, in similar environments, with a comprehensible and boundable error regime, does not necessarily follow the same dynamic. The work of Rabanser et al. documents precisely this gap. They do not establish an impossibility of future convergence, but they show that, in the current state of observed systems, capability progression does not yet allow governance to be treated as a mere by-product of scaling.[2]

The second reason is that not all relevant risks reside in the model taken in isolation. Gartner's AI TRiSM framework targets the governance, trustworthiness, robustness, data protection and security of AI deployments. When extended in the literature devoted to multi-agent and agentic systems, it illuminates classes of systemic risks that cannot be reduced to the intrinsic quality of a single component: error cascades, unforeseen interactions, circumvention of oversight mechanisms, or degradation induced by shared context and memory spaces. A locally satisfactory component can therefore produce globally ungoverned effects when inserted into a broader interaction system. The governance of these properties is necessarily systemic.[4]

The third reason is institutional. Emerging public frameworks do not wager on a future behavioral self-sufficiency of models. The Singaporean framework from IMDA, the NIST initiative on agent standards, the clarifications from the British regulator on the application of consumer law to commercial agents, and the European legal framework converge on a single

point: accountability, auditability, meaningful human oversight and traceability remain requirements of the system, not hopes placed in a component.[5][6][7][8]

The initial thesis must, however, be refined. To say that agentic governance is exogenous does not mean that the endogenous properties of the model are immaterial. They are not. A system is easier to govern if its components better respect expected formats, produce more verifiable justifications, manage their uncertainty more cleanly, or conform more closely to structured interfaces. Endogenous advances reduce the cost of governance; they do not replace it. The correct relationship is therefore not one of absolute opposition, but of hierarchy. Governability must be guaranteed by the architecture, even if it can be facilitated by internal advances of the components.

The problem is therefore not the agent per se. It is the admission regime for its decision within the system.

3. Action-Space, Autonomy, Reversibility: A Risk Grid for Reasoning About Decision Regimes

The principal conceptual contribution of the IMDA framework devoted to agentic AI is to shift risk assessment from the model alone toward the combination of several operational properties. Two of them are explicitly central: the extent of the action-space and the degree of autonomy. A third, essential in practice, must be integrated from the outset rather than appended as a footnote: reversibility. The IMDA document defines autonomy as the degree to which an agent can decide when and how to act toward a goal, and treats the reversibility of actions as a structuring variable for risk analysis.[5]

The action-space designates the perimeter of tools, data, systems and action surfaces to which the agent effectively has access. An agent limited to read-only consultation of a documentary base does not expose the same risk profile as an agent capable of modifying a CRM, reclassifying documents, sending a message to a client, validating a transaction or publishing content in an external environment. The action-space is not an abstract attribute. It defines the concrete scope of what can occur in the event of an erroneous, misleading, incomplete or opportunistic decision.

Autonomy designates here the degree of latitude effectively left to the system to select its means, chain its steps, arbitrate between options or trigger an action without explicit prior validation. Autonomy is not an essence of the agent. It is a property configured by the system: instruction level, permissions, presence of checkpoints, escalation conditions, ability or inability to call certain tools, and validation modalities before execution.

Reversibility qualifies the cancellable, correctable or opposable character of the action. An error in a meeting agenda suggestion does not carry the same status as an error in sending a regulatory notification, in the mass reclassification of a corpus, in the modification of a client file, or in triggering a financial order. Some actions are easy to reverse; others leave persistent

traces, produce cascading effects, or create legal or reputational consequences that are difficult to repair.

These three variables form a more robust grid than the simplistic opposition between "copilot" and "autonomous agent." They allow reasoning in terms of decision regimes.

The first regime is that of assistance. The agent proposes, reformulates, synthesizes, critiques, compares, prepares, but executes nothing beyond the production of an artifact interpretable by the human. Its action-space is purely informational or read-only, its autonomy is low and its effects are highly reversible. Governance is relatively light here, not because the system is intrinsically reliable, but because its impact space is strongly bounded. The major risk is not the technical autonomy of the system; it is the cognitive passivity of the user, that is, complacency or rubber-stamping.

The second regime is that of structured recommendation. The agent no longer merely produces free text; it pre-fills a decisional artifact, links its recommendations to sources, flags points of attention, highlights conflicts, prepares an arbitration or a documentary diagnosis. Its action-space can be broader, including multiple corpora, tickets, reference systems, code elements or business data. Its autonomy is intermediate, since it structures the problem and prefigures the decision without concluding it legally or operationally. In this regime, governance requires that each recommendation be traceable to a provenance, that the rules mobilized be identifiable, and that final responsibility remain clearly human. The major risk is no longer complacency alone; it is automation bias, the tendency to excessively credit the system's proposal simply because it is formally structured.

The third regime is that of bounded execution. Here, the agent effectively acts upon an environment, but within a closed perimeter, explicitly authorized, accompanied by verifiable invariants and an escalation mechanism whenever those invariants are no longer satisfied. It can, for instance, enrich metadata according to predefined taxonomies, trigger a standard workflow, classify documentary artifacts, generate a formal draft within a constrained template, or perform a limited and reversible write action in a controlled zone. The core of governance is then no longer solely the ex ante human validation of each action, but the combination of strict allowlists, declarative decision policies, periodic human sampling and robust evidence logs. The major risk here is silent drift: an apparently minor error, repeated at scale, can produce lasting systemic degradation without any immediately spectacular incident.

These regimes are not intended to form a moral hierarchy or an automatic trajectory toward greater autonomy. They are not maturity levels; they are distinct architectural modes. The same system can operate in assistance mode on certain tasks, in structured recommendation mode on others, and in bounded execution mode on a very restricted subset of reversible actions. What matters is that the regime be explicit, documented, revisable and justifiable.

Two modulators must furthermore be added to this grid to make it a genuinely operational framework. The first is domain criticality. A given action-space, autonomy and reversibility configuration does not carry the same weight depending on whether it applies to low-stakes internal classification, documentary compliance, an act with effects on rights, or a clinical, financial or legal environment. The second is error asymmetry. Some systems tolerate false positives or false negatives relatively well. Others do not. A serious governance architecture cannot ignore these damage distributions.

The consequence is clear. Agentic governance must not start from the question "how far can we let the agent go?" but from the question "in what decision regime can this artifact be admitted, under what conditions, with what scope of action, what reversibility and what escalation mode?" This reversal is the condition of any mature governance.

4. From Declarative Governance to Admission Control: Toward Decision Contracts

Once it is accepted that governability must be produced by the system, the question remains: what architectural form can give it operational consistency? The most fruitful analogy is not to be found in the psychological metaphors of the "responsible" or "aligned" agent, but in certain established practices of distributed software engineering. When a technical environment must govern the access of components to shared resources, it does not presume their virtue. It institutes mechanisms of admission control, authorization, declarative policy and audit.

The comparison with Kubernetes must be handled with precision. The point is not to claim that an agent and a container are of the same order. A container executes deterministic code on a given input; an LLM agent operates in a stochastic and partially interpretive space, where the same input can produce different action sequences. The analogy therefore bears neither on cognition nor on execution determinism, but on the governance structure: in both cases, one does not ask a component to be morally safe; one bounds what it can do, under what conditions and according to what declarative policies.

It is in this logic that the notion of the decision contract takes its meaning. It should not be presented as an already stabilized industry standard. It is not. More rigorously, it is a proposed architectural pattern derived from the design-by-contract logic formalized by Bertrand Meyer, which transposes the idea of preconditions, postconditions and invariants from the domain of software correctness to that of decisional governance.[9] The decision contract is not reduced to a post hoc log. It describes, prior to the action or state change, what the system intends to do, on what sources it relies, in what decision regime it operates, under what policies, with what rights, what invariants, what level of reversibility and what escalation conditions. This contract is then submitted to an admission mechanism, deterministic as far as possible, assisted if necessary, but distinct from the component that requests to act.

The value of this pattern is threefold. First, it decouples reliability from governability. An agent can remain imperfect, variable, even sometimes questionable in its intermediate reasoning, while being inserted into a system that prohibits it from acting outside the authorized framework. Governability no longer depends on whether the component "behaves well," but on whether the system filters what is admissible to transform into action. Second, it makes policies explicit, versioned and auditable. A rule such as "no irreversible action may be executed in bounded execution mode without nominative human validation" becomes a governance artifact, not a diffuse intention buried in documentation or a prompt. A policy such as "no non-allowlisted tool may be called from this context" becomes verifiable. The governance system thereby gains in opposability. Third, it transforms the trace. A simple posterior log records that an action took place. A well-designed decision contract allows reconstruction of why an action was proposed, with what sources, under what regime, with what applicable policy, in what context and with what validation. In regulated environments, this difference is decisive. What organizations must be able to produce is not merely the raw history of actions, but the admissibility chain that made those actions possible.

One must, however, avoid overselling this approach. Decision contracts, at this stage, belong more to a promising architectural pattern than to a mature industry standard. Existing implementations are still fragmentary, often experimental, sometimes carried by demonstrative projects rather than by documented large-scale deployments. It would therefore be excessive to speak of them as a stabilized industry achievement. Their interest lies precisely in their intermediate status: they offer a credible design language for transforming governance principles into system primitives.

Small-scale experiments nonetheless show that such a pattern is implementable. When a monitoring, scoping or drafting agent is not authorized to advance an artifact to the next stage without passing through a form, rigor or compliance check, the system is already practicing, at a certain scale, governance by admission. This does not prove the general validity of the pattern. But it indicates that it does not belong to pure abstraction.

The real difficulty is not conceptual. It is economic and technical. Introducing an admission control layer adds latency, complexity, a policy cost, a maintenance cost and often a human cost. A governed agentic system is more expensive to design, to evolve and to supervise than an agentic system left to itself. This additional cost is not a regrettable accident. It is the price of governability.

5. The Institutional Frameworks of 2026: Partial Convergence, Different Logics

The year 2026 sees the emergence of several institutional frameworks pertinent to thinking about agent governance, but these frameworks share neither the same status, nor the same normative force, nor the same granularity.

The framework proposed by IMDA in Singapore constitutes to date one of the most directly oriented formulations toward agentic systems. Its principal value is not so much to discover ex nihilo previously unknown principles as to organize them around operational variables adapted to the agentic domain, notably the scope of action, the degree of autonomy and the necessity of meaningful human control mechanisms. Its regime is that of structuring soft law: it recommends, it frames, it guides, but it does not sanction by itself. Its strength is to propose a design grammar. Its limitation is that it does not, as such, provide a coercive legal regime.[5]

The NIST initiative on agent standards pursues a different logic. It does not situate itself primarily on the terrain of legal constraint, but on that of interoperability, technical trust, security and the standardization of interfaces or mechanisms that will allow agentic systems to be deployed in more predictable environments. NIST explicitly presents it as an initiative intended to foster confident adoption, secure operation and fluid interoperability of agents in the digital ecosystem. This is normative infrastructure work in the broad sense, closer to the fabric of standards than to administrative policing. Its importance is real, but of a different order from that of binding regulation.[6]

The case of the European Union is different still. The European AI regulation introduces a binding legal regime based on a classification by risk level and usage categories. A frequent simplification must be corrected here. A system is not high-risk because it is "agentic," "autonomous" or "multi-step" as such. The relevant legal criterion is not the architectural form of the system alone, but its potential inscription within the categories or contexts provided for by the regulation, notably the cases listed in Annex III, or its integration into certain regulated products. The European Commission recalls that Annex III lists high-risk use cases and that these categories are subject to specific framing.[8]

It follows that an autonomous agent may not fall under the high-risk regime, while a less spectacular system deployed in a legally sensitive context may fall fully within it. Agenticity is therefore not an autonomous legal criterion for qualification. This precision does not weaken the article's thesis. It makes it more robust. For even without abusively assimilating all autonomous agents to high-risk systems, the European framework supports the idea that a system deployed in sensitive uses, or producing significant effects on rights, obligations or access, will be more likely to enter regimes requiring human oversight, traceability, documentation and accountability.[8]

The United Kingdom, through the Competition and Markets Authority's guidance on the use of AI agents under consumer law, adds a useful clarification. The agent does not open a lawless zone. It constitutes a mode of execution or intermediation that remains subject to existing consumer law obligations, and the CMA explicitly recalls that failures may expose companies to enforcement measures and sanctions.[7]

Taken together, these frameworks do not converge toward a unified global law of the agentic. They converge toward a more modest but more important idea: no serious institutional actor

treats agentic governance as a simple question of internal model quality. All of them treat it, in different forms, as a system problem.

6. Limitations of the Thesis and Research Program

The thesis defended here is neither closed nor complete. It rests on a set of plausible inferences, supported by emerging work, institutional frameworks and engineering analogies, but it should not be presented as a completed empirical demonstration.

The first limitation is the absence of robust longitudinal data on the relationship between scaling, alignment and agentic reliability. Recent work documents a current empirical gap between capability and reliability on certain benchmarks and certain model families. It does not establish a theoretical impossibility of reducing this gap in the medium term. The thesis of strong exogenous governance must therefore be formulated as a prudent and architectural requirement in the present state of systems, not as a definitive refutation of any future partial convergence.[2]

The second limitation is the insufficiency of documented industrial cases. The patterns proposed here, notably the combination of decision regimes, decision contracts and admission control, are intellectually coherent, but still weakly supported by comparative publications showing their measured impact on incident frequency, compliance burden, operational velocity or total cost of ownership. The passage from the prescriptive to the empirical remains to be accomplished.

The third limitation is economic. A serious governance architecture has a cost. It adds friction, operational debt, policy layers, escalation mechanisms, documentary requirements, audit operations and often non-trivial human expenditures. As long as no solid literature systematically compares this cost to that of non-governance across different classes of systems, the argument will remain partially prudential.

The fourth limitation relates to the domain of validity of the proposed framework. The action-space, autonomy and reversibility grid works well for relatively bounded contemporary enterprise systems. It will be more difficult to apply to self-modifying systems, to agents crossing multiple organizational trust boundaries, to multi-agent environments exhibiting emergent behaviors not easily deducible, or to architectures where permissions themselves can be dynamically reconfigured by the system. The framework is not thereby invalidated; it is situated.

The fifth limitation is geographical and documentary. The sources mobilized in this field remain very largely Anglophone, notably complemented by Singapore. Non-Anglophone Asian frameworks are underrepresented in the current mapping: the guidelines of the Japanese AI Strategy Council, the Chinese regulatory framework jointly led by MIIT and the CAC, and Korean preparatory work toward a national AI Act all constitute pertinent sources

that the present analysis does not mobilize. This asymmetry limits the claim to exhaustiveness and constitutes in itself a methodological bias to be declared.

The sixth limitation concerns the fragility of governance itself. The article posits that a well-designed exogenous governance architecture can compensate for the shortcomings of models. But this architecture is itself a constructed, maintained and operated artifact by organizations that are not immune to their own failures. Policies that are too loose because calibrated under operational pressure, allowlists that expand by accretion without periodic review, escalation mechanisms systematically bypassed by habit, decision contracts that are formally correct but substantively empty, all constitute failure modes of governance by governance. The quality of an admission control system depends not only on its technical primitives; it depends on the organizational discipline that maintains them. This recursion, "who governs the governance?" does not invalidate the approach, but it forbids presenting it as self-sufficient. A governance architecture is a necessary condition, not a guarantee.

These limitations do not annul the thesis. They rather define a research program. The first question is empirical: to what extent does the gap between capability and reliability effectively narrow from one model generation to the next, once changes in benchmark, context and evaluation protocol are controlled for? The second is architectural: do the patterns of admission control, declarative policy and decision contracts transpose effectively to the stochastic domain without producing prohibitive complexity? The third is economic: above what threshold of criticality, usage frequency and error asymmetry does exogenous governance become rational relative to its total cost? The fourth is systemic: how can one govern a multi-agent ensemble when the global behavior of the system is no longer inferable from the local properties of each of its components?

A discipline of agentic governance will not emerge from a single paper, still less from a single taxonomy. It will emerge from the articulation between engineering, law, systems theory, compliance economics and industrial feedback.

7. The Price of Governability

The conclusion can now be formulated with greater precision.

The right question is not: how can we make agents sufficiently reliable that we can finally trust them without reservation? Posed in this way, the question perpetuates a tenacious illusion, that the central problem is psychological or moral, as if the system should ultimately deserve quasi-personal trust. The pertinent question is rather: how can we architect a system in which no significant decision accesses execution without having traversed an admissibility regime proportionate to its scope of action, its level of autonomy, its reversibility, its criticality and the asymmetry of its possible errors?

Trust, in this perspective, is not a heroic property of the model. It is a constructed property of the system. It does not suppose that the component is perfect. It supposes that what it can

transform into action is bounded, qualified, logged, controlled, revisable and attributable. It further supposes, in high-risk systems, that certain classes of decisions remain structurally unavailable to full autonomy and remain conditioned on an explicit human guarantee, not as a psychological concession to prudence, but as a principle for organizing accountability.

This thesis is not comfortable. It renounces one of the most seductive promises of the contemporary agentic imaginary, that of an increasing autonomy that would progressively eliminate the human bottleneck. This renunciation is, however, less a conservatism than a clarification. In high-consequence systems, the human bottleneck is not a transitional defect destined to disappear with model improvement. It constitutes an institutional guarantee. In domains where a decision can engage collective safety, the integrity of critical infrastructure, the conduct of a military operation, nuclear exposure, or a patient's health status, no serious architecture should authorize a system to make the final decision alone without meaningful human oversight. What can become autonomous, under conditions, is therefore not the sovereign decision in the full sense. It is the bounded execution of constraints, sequences or operations already legitimated, under a control regime where human intervention remains the ultimate lock of admissibility.

The cost of this governability is real. Architectural complexity, latency, policy maintenance, human validation, documentary discipline, organizational friction. This cost is not a secondary defect. It is the exact price of a system capable of accounting for what it did, why it did it, within what framework it did it, and who bore the responsibility.

Models will continue to improve. It would be absurd to bet against that. But even better models will not eliminate the need for a governance architecture. At best, they will reduce its marginal cost.

An agent is not governed because it is better. It is governed because a system prohibits it from acting outside an explicitly authorized decision regime.

In high-risk systems, autonomy must not eliminate the human bottleneck. It must learn to live under it.

Notes

[1] Gartner, "Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026, Up from Less Than 5% in 2025," press release, August 26, 2025. https://www.gartner.com/en/newsroom/press-releases/2025-08-26-gartner-predicts40-percent-of-enterprise-apps-will-feature-task-specific-ai-agents-by-2026

[2] Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan, "Towards a Science of AI Agent Reliability," arXiv:2602.16666, February 18, 2026. https://arxiv.org/abs/2602.16666

[3] Simon Willison, "Writing about Agentic Engineering Patterns," February 23, 2026; see also Agentic Engineering Patterns, ongoing guide, simonwillison.net. References mobilized as practitioner

observations, not as market measurement. https://simonwillison.net/guides/agentic-engineering-patterns/

[4] Gartner presents AI TRiSM as a general framework intended to ensure governance, trustworthiness, fairness, reliability, robustness, efficacy and data protection in AI deployments. The present discussion extends its logic to the case of agentic systems in a doctrinal and systemic sense; the point is not to attribute to Gartner a detailed agentic formalization that it does not explicitly claim in this source. See Gartner, "Tackling Trust, Risk and Security in AI Models," December 24, 2024.

[5] Infocomm Media Development Authority, Model AI Governance Framework for Agentic AI, version 1.0, January 22, 2026, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf.

[6] National Institute of Standards and Technology, Center for AI Standards and Innovation, "Announcing the AI Agent Standards Initiative: Interoperable and Secure," February 17, 2026. https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure

[7] UK Competition and Markets Authority, "Complying with Consumer Law When Using AI Agents," March 9, 2026. https://www.gov.uk/government/publications/complying-with-consumer-law-when-using-ai-agents

[8] European Commission, "Navigating the AI Act," FAQ on Regulation (EU) 2024/1689, notably on high-risk systems and Annex III. https://digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act

[9] Bertrand Meyer, "Applying Design by Contract," Computer 25, no. 10 (1992): 40–51. https://doi.org/10.1109/2.161279

[10] The evaluations published by laboratories show inter-generational gains on certain dimensions such as instruction-following, output structuring and constrained format management. See for instance Anthropic, System Card: Claude Opus 4 & Claude Sonnet 4 (May 2025, updated July 2025), as well as OpenAI, GPT-4o System Card (August 8, 2024). These advances should not, however, be conflated with equivalent gains in consistency or agentic reliability in the sense of [2], nor with the possibility of removing human guarantees in high-consequence systems. In high-risk environments, endogenous model advances may reduce the supervision burden, but they cannot abolish the requirement for meaningful human oversight when decisions engage safety, the integrity of critical infrastructure, or substantial effects on persons.

Bibliography

Anthropic. "Model system cards." System cards index page, accessed 2026. https://www.anthropic.com/system-cards.

Anthropic. System Card: Claude Opus 4 & Claude Sonnet 4. May 2025, updated July 16, 2025. https://www.anthropic.com/claude-4-system-card.

European Commission. "Navigating the AI Act." FAQ on Regulation (EU) 2024/1689. https://digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act.

Competition and Markets Authority. Complying with Consumer Law When Using AI Agents. GOV.UK, March 9, 2026. https://www.gov.uk/government/publications/complying-with-consumer-law-when-using-ai-agents.

Gartner. "Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026, Up from Less Than 5% in 2025." Press release, August 26, 2025.

Gartner. "Tackling Trust, Risk and Security in AI Models." December 24, 2024. https://www.gartner.com/en/articles/ai-trust-and-ai-risk.

Infocomm Media Development Authority. Model AI Governance Framework for Agentic AI. Version 1.0. Singapore, January 22, 2026. https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf.

Meyer, Bertrand. "Applying Design by Contract." Computer 25, no. 10 (1992): 40–51. https://doi.org/10.1109/2.161279.

National Institute of Standards and Technology, Center for AI Standards and Innovation. "Announcing the AI Agent Standards Initiative: Interoperable and Secure." February 17, 2026. https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure.

OpenAI. "GPT-4o System Card." August 8, 2024. https://openai.com/index/gpt-4o-system-card/.

Rabanser, Stephan, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. "Towards a Science of AI Agent Reliability." arXiv:2602.16666, February 18, 2026. https://arxiv.org/abs/2602.16666.

Willison, Simon. "Writing about Agentic Engineering Patterns." February 23, 2026. https://simonwillison.net/2026/Feb/23/agentic-engineering-patterns/.

Willison, Simon. Agentic Engineering Patterns. Ongoing guide, 2026. https://simonwillison.net/guides/agentic-engineering-patterns/.