THE AGENTIC READINESS GATE

A diagnostic of your organization's readiness to deploy and govern agentic AI, scored on evidence, one use case at a time

Method note, version 2.6

Vivek Hans, Genitive.AI

DOI: https://doi.org/10.5281/zenodo.23021021

1. What it is

The Agentic Readiness Gate scores how ready an organization is to run one agentic AI use case: one agent, or several agents doing one job together. It asks eight readiness checks. Six are control areas, where a gap can hurt someone. Two are value areas, where a gap gets the project cancelled. Each area is scored 1 to 5 against written descriptions at levels 1, 3 and 5, and every score points to evidence.

The output is a ceiling and two actions. The ceiling is the lowest control score, never an average. It gates the decision about this use case: it says what not to do yet. It never certifies that a deployment is safe. Each action has an owner, a date and the evidence that shows it is done.

The instrument is a directional self-assessment on paper, designed for a working session. It is not a validated audit, and it is not a maturity model for the whole organization.

It applies to agentic AI in any industry. Its first public use is with life sciences data teams at AnalytiCon 2026 (13 October 2026), and the worked case in section 7 comes from that setting. Its use elsewhere rests on the cross-industry guidance it draws on (section 8), not yet on use in other sectors.

2. Why agentic readiness differs from AI readiness

AI in the ordinary sense produces outputs that a person acts on: a forecast, a risk score, a drafted summary. A person stands between every output and every action, so governance can centre on the output. An agent carries out tasks on its own, within limits someone set: it runs queries, calls tools, changes records and picks its next step from what it finds. Governance then has to cover the system that acts: what it may do alone, whose identity it acts under, how it is stopped and undone, and how anyone would know it had stopped working.

The eight areas follow from that shift. Output becoming action forces identity and accountability (areas 6 and 4). Acting on retrieved or remembered facts forces provenance (area 5). Chained steps force testing and monitoring (area 7). Deciding in advance what the agent may do alone, and being able to stop and undo it, forces areas 2 and 3. Ongoing operation forces cost and drift (areas 7 and 8). Area 1 does not come from the shift, since any AI needs a purpose and an owner; it stays because area 8 measures impact against its target, and because unclear business value is one of the reasons Gartner gives for cancelled agentic projects.

Workforce, culture, strategy and general data quality are deliberately excluded. They matter, but they are AI readiness, and general indices already measure them. An organization can be AI-mature and still not ready for a particular agent.

3. The eight checks

Area	Job	Question
1 Business outcome	Value	How clearly is its purpose defined, owned and measured?
2 Risk appetite	Control	How firmly are the actions it may take on its own set, and enforced by the system?

Area	Job	Question
3 Workflow design and recovery	Control	How well are its handoffs and its recovery designed around how it fails?
4 Human-in-the-lead	Control	How reliably would the final human check catch an error?
5 Data provenance	Control	How easily can you trace what it used, including what it remembers?
6 Agent identity and entitlements	Control	How tightly is its access scoped, and how fully are its actions traced to it?
7 Evaluation and monitoring	Control	How well can you show it works, and how quickly would you know if it stopped?
8 Unit economics and impact	Value	How well do you know what each accepted case costs, and what it delivers?

Control areas (2 to 7) set the ceiling. Value areas (1 and 8) decide whether to scale.

4. Level descriptions

Only levels 1, 3 and 5 are described. Level 3 is written as separate statements so that "some" and "all" can be counted. The wording below is identical to the printed scorecard.

Area	Level 1	Level 3	Level 5
1 Business outcome	Exists because agents matter. No number named, no owner.	Named owner and target, usually cost or cycle time, with a baseline measured before build.	The outcome is sustained and evidenced over time, whatever the aim, and owned like any operating commitment.
2 Risk appetite	No stated position. Whoever builds it decides what it may do alone.	A written rule, by type of action (read, draft, change, irreversible), for which actions need a person's sign-off, applied to every new use case. Enforced through permissions and approval steps, not only instructions to the agent.	Tiered by reversibility, with standing approval only for routine actions, reviewed as the agent's track record and the models change, and teams can cite it unprompted.
3 Workflow design and recovery	Agent added to the existing process. Checkpoints sit where the old steps were, not where this agent makes mistakes. No agreed way to stop it or undo what it did.	Handoffs documented, including when the agent must stop and pass to a person. A documented way to stop it and reverse or contain what it did.	Checkpoints placed where this agent's errors actually occur, based on its track record. Stopping it, rolling back its actions and revoking its access are rehearsed.
4 Human-in-the-lead	Oversight is a sign-off box. The signer lacks the time and the source material.	Named qualified reviewers who see the agent's sources and trail, with authority to stop the workflow.	Oversight itself is measured: override and edit rates tracked, known errors planted in controlled tests of reviewers, never in live records. Depth matches consequence.
5 Data provenance	Provenance reconstructed by hand when someone asks, if at all.	Retrieved statements carry their source and version. Reachable sources are governed and documented.	Provenance automatic end to end, including which version applied at the time. Retrieval scoped so it cannot surface what it should not. What it keeps between tasks is governed like any other source.

Area	Level 1	Level 3	Level 5
6 Agent identity and entitlements	Runs under a shared account or a person's full credentials. Nobody could separate its actions from a person's.	Own identity, or a person's access narrowed to the task and marked as the agent's. Actions logged well enough to reconstruct. The tools, connectors and other agents it can call are approved and listed. Any code it writes runs in isolation. Calls to and from other agents are authenticated.	In access review and decommissioning like any privileged identity, with a named owner. If it reads outside content, it cannot send or change data without a check.
7 Evaluation and monitoring	You would find out from a complaint or an audit finding.	Before go-live, a test set of your own cases with known right answers and agreed pass thresholds, including cases built to mislead it. In production, runs traced and scored against them.	Any change to model, prompt, tools or permissions, data or memory, or the process it serves reruns the test set before release. Production failures are added to it. Changes you did not make are detected.
8 Unit economics and impact	Neither. Cost surfaces on an invoice. Value is asserted, not measured.	Cost per accepted case tracked, including review, corrections, retries and support, with spend and step limits, and impact measured against the area 1 target.	Unit economics shape design before build: which model, how many steps, where a person reviews. Impact measured consistently enough to compare use cases.

5. Scoring rules

- **First question.** Does it act on its own along the way: run a query, call a tool, change a record? This counts even if a person reviews the final result. If it only suggests and a person acts on every output, it is AI, not an agent, and the card is not for it. If it acts, who picks the next step: steps someone wrote (a fixed workflow, still scored) or the system itself, from what it finds (an agent)? Where steps are known and inputs structured, code or rules will usually do the job at lower cost.
- **Scope.** Score one use case as written: what the agent does and what a person decides. If several agents do the job, score the system; each check must hold at every agent and every handoff, and the weakest sets the score. The orchestrator is part of the system, not a separate use case.
- **Ladder.** Tick the level 3 statements you can show: none is 1, some is 2, all is 3. All of level 3 plus some of level 5 is 4. All of level 3 and all of level 5 is 5. Torn between two scores, take the lower.
- **Cannot apply.** A statement that cannot apply to the design (it writes no code; it calls no other agent) counts as shown, with the reason written down. It never covers a statement that is merely inconvenient.
- **Unknowns.** Mark U. A U counts as 1 for the ceiling, written "1 (unverified)", and reads as below 3 in a value area. Write who would know. Three or more U’s make finding out the first action.
- **Today, not the plan.** Score what is in place today for the design you would deploy. Planned controls go beside the score, not in it.
- **Ceiling.** The lowest score in areas 2 to 7. It holds for the use case as written. Scaling beyond the scope that was evidenced (for example, from a pilot group to every user) goes in stages, checking reviewer capacity, access and cost at each step.
- **Value reading.** If area 1 or area 8 is below 3, or U, do not scale, even if the ceiling allows it. A value reading never permits a pilot on its own; a pilot still needs its controls.

- **Two actions.** Start with the two lowest areas; at least one must be a control area. Ties go first to the area that blocks what the agent will do; if still tied, use the order 1, 5, 2, 3, 6, 4, 7, 8, which is a convention. Each action targets the next described level (3 from 1 or 2, 5 from 3 or 4) and has an owner, a date and evidence you would see, not be told. One action may be swapped for a higher-scoring area that blocks what the agent will do, with the reason written down.
- **Calibration.** Score privately first. Discuss patterns, never individual scores. A second scorer from the same organization is the cheapest correction. A one-level difference can be noise.

6. Why the ceiling is the lowest control score

The ceiling is a gating rule, not a calculation. It works like a required check in a software release pipeline: one failed check blocks the release, however many others pass. In pharmaceutical manufacturing the same logic is batch release: nine release tests pass, one is out of specification, and the batch does not ship. Nobody averages the results. A strong score in one control area does not compensate for a borrowed login in another, because the failure that reaches a customer, a patient, a regulator or a production system comes through the weakest control.

Non-compensatory rating is established within capability domains in CMMI, C2M2 and ISO/IEC 33020, where "little or no evidence of achievement" means the attribute is not achieved. The Gate applies the same logic across the six control areas of one use case. None of the 22 frameworks reviewed in September 2026 caps readiness at the weakest control; most average or weight their dimensions.

Because the ceiling is the lowest score, fixing it moves the ceiling only to the next lowest area. That is intended. It points each round of work at the current constraint and shows why a re-score, not a one-off assessment, is the unit of progress.

7. Worked case: one agent, two teams (a life sciences example)

The case is fictional. An analytics request agent answers data questions from medical affairs and commercial colleagues. It writes a query, runs it against the data warehouse, decides whether it needs another, and drafts the answer. An analyst reviews every draft before it goes back. It works alone and calls no other agent. At two companies it has run for three months with ten people, and both want to open it to every medical affairs and commercial colleague. It is an agent even though an analyst reviews every answer, because it runs its own queries and picks the next one.

Area	Team A	Team B	A	B
1 Business outcome	The head of medical analytics owns it. Target: routine requests answered in one day instead of five. Baseline measured before build; results reported monthly to leadership.	Owner and target named. Baseline measured before build.	4	3
2 Risk appetite	A written rule, by type of action: it may read and draft; nothing leaves analytics without an analyst's approval. It applies to every agent, and the query tool rejects anything but reads.	The same written rule, enforced the same way.	3	3
3 Workflow design and recovery	A documented rule hands an ambiguous question to an analyst. An analyst can hold an answer, but there is no agreed way to switch the agent off or recall an answer already sent.	Documented handoffs: it stops and passes to an analyst when the question is ambiguous or data is missing. A documented procedure lets the team switch it off and recall any draft; analysts answer by hand meanwhile.	2	3

Area	Team A	Team B	A	B
4 Human-in-the-lead	A named analyst, trained on the data marts and given time for every review, checks each answer with the query, the tables and the numbers in view, and has authority to stop the workflow. How often analysts correct answers is tracked monthly.	A named analyst, trained on the data marts and given time for every review, checks each answer with the query and tables in view, and has authority to stop the workflow.	4	3
5 Data provenance	Every number links to its query, table and data version, automatically, including the version on the date asked. Only certified, documented data marts are reachable.	Every number links to its query, table and data version. Only approved, documented data marts are reachable.	4	3
6 Agent identity and entitlements	It runs on the head of analytics' own login, which can read and change every table in the warehouse. The logs show her name. Nobody has listed the tools it uses, and code it writes runs directly on the warehouse server.	Its own service identity, read access to approved data marts only, every query logged as the agent's. Its tools are approved and listed; any code it writes runs in a sandbox.	1	3
7 Evaluation and monitoring	Tested on past requests with known answers and pass thresholds agreed first, including trick questions. Production answers scored against them; any model or prompt change reruns the tests.	Tested on past requests with known answers and pass thresholds agreed first, including trick questions. Production answers scored against them.	4	3
8 Unit economics and impact	Cost per accepted answer tracked, including review, corrections, retries and support, with spend and step limits, against the one-day target. Cost per answer decided which model it uses.	Cost per accepted answer tracked, including review, corrections, retries and support, with spend and step limits, against the one-day target.	4	3

Team A averages 3.3 with a ceiling of 1. Team B averages 3.0 with a ceiling of 3. Team A looks stronger on most rows and runs on a manager's full login; its identity score of 1 sets the ceiling. The call: Team A waits until the agent has its own identity and an agreed way to switch it off and recall an answer. Team B opens up in stages, checking reviewer capacity and cost per answer at each step. When Team A fixes identity, its ceiling moves to 2, because recovery (area 3) is now the lowest.

8. Lineage and prior work

The scoring discipline was informed by the AI Adoption Maturity Model released by Carnegie Mellon University's Software Engineering Institute and Accenture in June 2026. The Gate applies that discipline to a single agent use case rather than an organization's AI programme.

The NIST AI Risk Management Framework (2023) describes use-case profiles: implementations of the framework for a specific setting or application, based on its requirements, risk tolerance and resources. The Gate's one-use-case scope follows the same idea for agentic AI in any sector.

The ISPE D/A/CH "AI Maturity Model for GxP Application" (Erdmann et al., 2022) is the closest life sciences precedent. It classifies AI applications on two dimensions, control design and autonomy, to set validation activities. The Gate shares its premise that controls must rise with autonomy, and differs in scoring organizational readiness for one agent use case, with a weakest-control ceiling.

The control areas converge with agent-specific guidance published in 2025 and 2026: Singapore IMDA's Model AI Governance Framework for Agentic AI, the Cloud Security Alliance's draft agentic governance maturity model and MCP guidance, the AWS Agentic AI Security Scoping Matrix, the OWASP Top 10 for Agentic Applications, joint Five Eyes guidance, and OpenAI's practices for governing agentic AI systems. They agree on bounded actions, agent identity, checkpoints, adversarial testing and recovery. The Gate's contribution is to score those controls for one use case, on evidence, with the weakest control as the ceiling.

The World Economic Forum's readiness framework for government (April 2026) classifies government functions by their agentic AI potential and their implementation complexity. It helps decide where to start. The Gate answers the next question: whether an organization is ready to run the use case it has chosen.

General AI maturity and readiness models, such as Gartner's AI Maturity Model and Cisco's AI Readiness Index, score an organization's whole AI programme across stages or weighted pillars. They remain useful. They answer a different question.

In life sciences, the FDA and EMA joint guiding principles for good AI practice in drug development (January 2026) ask for a well-defined context of use, which the one-use-case scope also meets.

9. Limitations

- **Not validated.** This is the instrument's first public use. Its reliability and its relationship to outcomes have not been tested. Maturity models as a class have a documented validation gap (Wendler, 2012).
- **Assessor error.** In a study of practitioners rating a described company against defined maturity levels, ratings deviated from the reference by about one level on average (mean 1.07; Schmitz et al., 2021). Self-assessment may add further bias. Treat a one-level difference as possible noise.
- **Evidence thresholds vary.** How much evidence is enough depends on the use case, its stage and what the agent may do. The workshop's tier guidance is explanatory and not part of the scoring.
- **The ceiling is a gate, not a guarantee.** A high ceiling means no control area blocks the use case as scored. It does not establish capacity, performance or cost at a larger scale.
- **Scope.** The Gate does not measure workforce readiness, culture or strategy, and it does not replace validation, security review or regulatory assessment.

10. Version history

v1 (August 2026) and v2 (September 2026): eight areas derived from the AI versus agent distinction. v2.2: control and value split; two actions rule; U flag. v2.3: identical question wording across card and deck; recovery, memory, tool and agent supply chain, and adversarial testing added to the descriptions. v2.4: level 4 defined as all of level 3 plus some of level 5. v2.5: one agent or several scored as one use case. v2.6 (29 September 2026): the first question asks whether it acts on its own along the way; level 5 includes all of level 3; statements that cannot apply count as shown with a reason; authenticated calls between agents move to level 3; planted errors only in controlled tests; score today and record plans separately; the ceiling is a gate that never certifies safety; staged scaling; case evidence stated in full.

11. How to cite

Hans, V. (2026). *The Agentic Readiness Gate: A diagnostic of your organization's readiness to deploy and govern agentic AI, scored on evidence, one use case at a time* (Version 2.6) [Method note and scorecard]. Zenodo. https://doi.org/10.5281/zenodo.23021021

Web: genitive.ai/agentic-readiness-gate. Drafting was assisted by AI tools; the author reviewed all content and is responsible for it.

References

AWS (2025). The Agentic AI Security Scoping Matrix. AWS Security Blog, 21 November 2025. https://aws.amazon.com/blogs/security/the-agentic-ai-security-scoping-matrix-a-framework-for-securing-autonomous-ai-systems

Cemri, M., et al. (2025). Why Do Multi-Agent LLM Systems Fail? arXiv 2503.13657. https://arxiv.org/abs/2503.13657

Cisco. AI Readiness Index methodology. https://www.cisco.com/c/m/en_us/solutions/ai/readiness-index/methodology.html

Cloud Security Alliance (2026). Agentic AI Governance Maturity Model, draft v1. https://labs.cloudsecurityalliance.org/agentic/agentic-governance-maturity-model-v1/

EMA and FDA (2026). Guiding principles of good AI practice in drug development, 14 January 2026. https://www.ema.europa.eu/en/documents/other/guiding-principles-good-ai-practice-drug-development_en.pdf

Erdmann, N., Blumenthal, R., Baumann, I., Kaufmann, M. (2022). AI Maturity Model for GxP Application: A Foundation for AI Validation. Pharmaceutical Engineering, March/April 2022. https://ispe.org/pharmaceutical-engineering/march-april-2022/ai-maturity-model-gxp-application-foundation-ai

Gartner. AI Maturity Model and AI Roadmap Toolkit. https://www.gartner.com/en/chief-information-officer/research/ai-maturity-model-toolkit

IMDA (2026). Model AI Governance Framework for Agentic AI. https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf

NIST (2023). AI Risk Management Framework 1.0, section 6, AI RMF Profiles. https://airc.nist.gov/airmf-resources/airmf/6-sec-profile/

OpenAI (2023). Practices for Governing Agentic AI Systems. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf

OWASP GenAI Security Project (2025). OWASP Top 10 for Agentic Applications for 2026, 9 December 2025. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

Schmitz, Schmid, Harborth and Pape (2021). Maturity level assessments of information security controls: An empirical analysis of practitioners' assessment capabilities. Computers & Security. https://www.sciencedirect.com/science/article/abs/pii/S0167404821001309

Software Engineering Institute, Carnegie Mellon University, and Accenture (2026). AI Adoption Maturity Model, 8 June 2026. https://www.sei.cmu.edu/news/sei-and-accenture-release-ai-adoption-maturity-model-to-help-organizations-scale-ai-with-predictable-outcomes/

World Economic Forum (2026). Making Agentic AI Work for Government: A Readiness Framework, 24 April 2026. https://www.weforum.org/publications/making-agentic-ai-work-for-government-a-readiness-framework/

Wendler, R. (2012). The maturity of maturity model research: A systematic mapping study. Information and Software Technology 54(12). https://www.sciencedirect.com/science/article/abs/pii/S0950584912001334