

Insight Brief

Validating AI Agents and Software

Building defensible trust as agentic systems reshape how software is developed, validated, and governed in healthcare and life sciences

LUK ARBUCKLE, Global AI Practice Leader in Defensible Data & AI, AI & Technology Solutions, IQVIA

GARETH DABBS, Vice President of Enterprise Platforms, AI & Technology Solutions, IQVIA



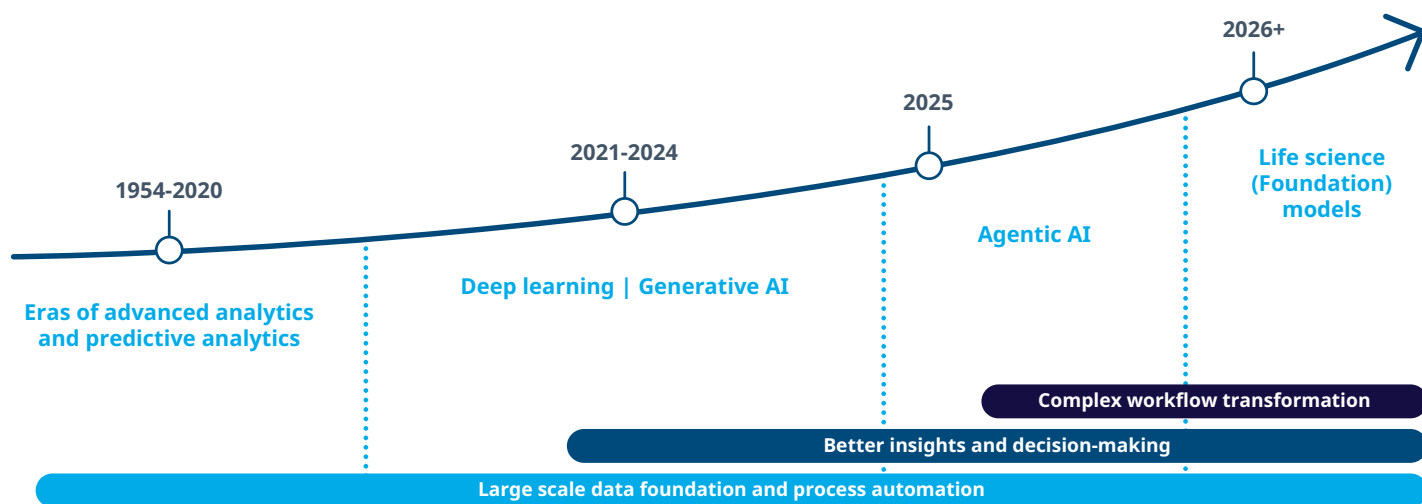
Table of contents

Trust is the gating question in the agentic era	1
The bar has moved from trustworthy to provable	2
From responsible and trustworthy to defensible AI	3
Proportionate risk tiering as a defining principle	4
Autonomy is the amplifier that agents introduce	4
Scientific validation as proportionate evidence	6
A staged view of the principles in action	6
Approaches to trust using evidence as by-product	7
Audit logs and traceability	7
User interfaces as trust surfaces	7
Control by design through guardrails, policy-as-code, and tool-gating	8
Using agents to validate agents	8
Lifecycle governance and continuous assurance	10
Conclusion	11
References	12
About the authors	15
About IQVIA	15

Trust is the gating question in the agentic era

Software is being built in a fundamentally different way. A decade ago, code was written line by line by engineers, reviewed by peers, and tested against specifications a human could read and reason about. Today a growing share of software is drafted, refactored, and tested by AI assistants, and an emerging class of systems, called

agents, no longer merely produce code or predictions but take actions. They call tools, query databases, draft and send communications, update records, and orchestrate multi-step workflows with limited human intervention. This is a qualitative shift that changes the central governance question from whether a prediction is accurate to whether an action is authorized, appropriate, and reversible.



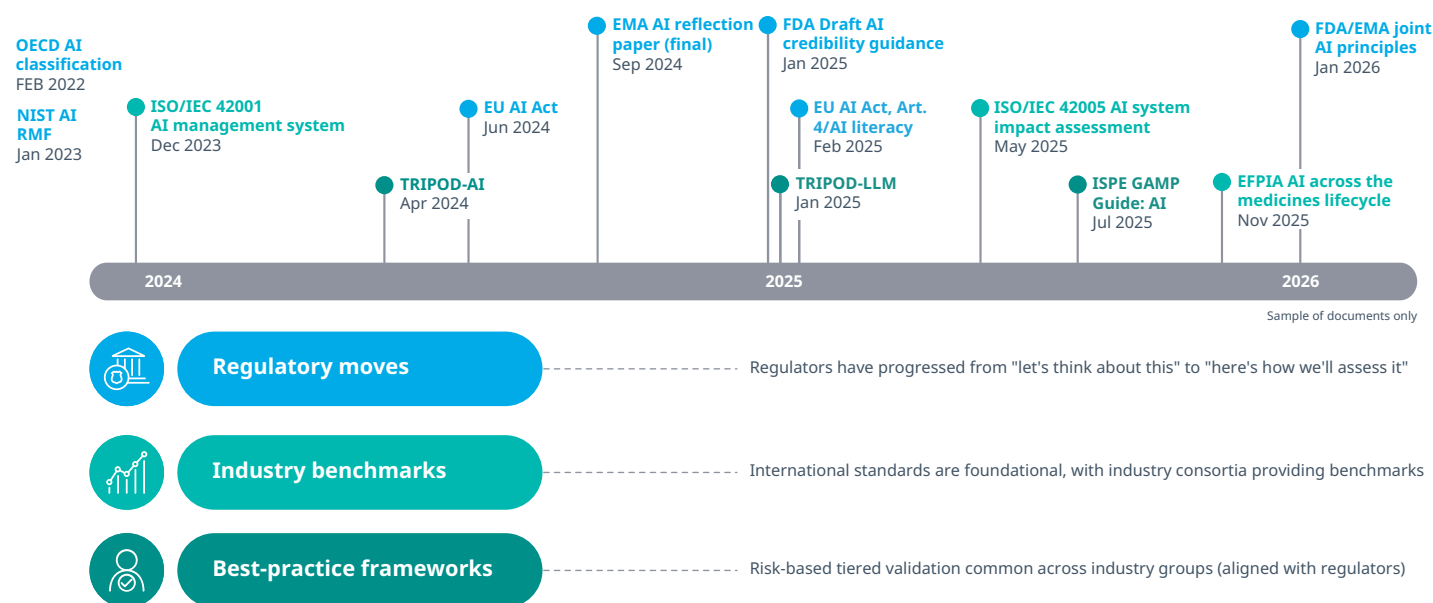
The properties that make agents powerful are the same ones that make them hard to govern. Their behavior emerges at runtime rather than being fixed at design time, so testing and code review capture only a fraction of the risk surface. Their reasoning is probabilistic, their execution path is dynamic, and they can delegate to other agents, creating accountability chains that traditional software governance never anticipated. In healthcare and life sciences, where an action can influence a regulated decision, a patient interaction, or a product-quality release, the tolerance for unexplained behavior is low and the expectation of evidence is high.

The question our customers are asking has therefore converged on two words that recur across very different conversations: validation and trust. They want to know that AI-assisted software does what it's supposed to do, that an agent will act only within its remit, and that when it does act, someone can reconstruct what happened and why. Interest in this question is expected in clinical settings. But interest is just as strong for the operational and commercial software that agents are increasingly asked to build and run.

The bar has moved from trustworthy to provable

The regulatory ground has shifted decisively in roughly twenty-four months, moving from “let us think about this” to “here is how we will assess it.” On the medicines side, the EMA’s reflection paper on artificial intelligence in the medicinal product lifecycle was finalized in September 2024, the FDA’s January 2025 draft guidance set out a seven-step, risk-based credibility framework

with lifecycle maintenance, and the two agencies have since converged on a shared set of AI principles.^{1, 2, 3} International standards now provide the foundation, with ISO/IEC 42001 AI management system and ISO/IEC 42005 AI system impact assessment, while industry and reporting standards such as TRIPOD+AI, TRIPOD-LLM, ISPE’s GAMP guidance, and EFPIA’s work on AI across the medicines lifecycle formalize risk-based, tiered validation.^{4, 5, 6, 7, 8, 9}



Alongside product-level expectations, a quieter obligation has reframed trust as a workforce question. Article 4 of the EU AI Act, the AI literacy provision, has applied since February 2025 and reaches every provider and deployer of AI systems regardless of risk tier.¹⁰ It requires organizations to take measures so that staff, and other people acting on their behalf, have a level of understanding proportionate to their role, the use case, and the people affected.¹¹ Gartner suggests AI literacy is no longer an HR initiative but a market-access requirement, and organizations that can’t demonstrate readiness will risk losing opportunities before AI value is ever realized.^{12, 13} Human readiness has become part of AI value, and part of the evidence that a system can be trusted.

Read together, these developments point to a single conclusion. Organizations are increasingly asked whether their AI is trustworthy, whether they can

prove it, and whether the people building and using it understand it. That’s a higher bar than a one-time review, and it’s the bar this paper is written to address.

We set out the principles by which agents and AI-assisted software can earn adoption and withstand scrutiny. It’s deliberately a discussion of principles rather than a procedures manual. We outline a defensible operating model, the logic of proportionate risk tiering, the essentials of scientific validation, and the engineering approaches, audit logs, interfaces, guardrails, and the use of agents to validate agents, that turn trust from an assertion into an artefact. The detailed evidence standards that operationalize these principles exist separately and will evolve as the regulatory landscape adapts.

From responsible and trustworthy to defensible AI

It helps to be precise about words that are often used interchangeably. Responsible AI states principles and is usually policy focused. Trustworthy AI tends to be more technical, concerned with properties such as reliability, robustness, and fairness. *Defensible AI* is what produces the evidence that a system can withstand scrutiny throughout its lifecycle, beyond merely pass a review once. It's the operating standard that turns good intentions and good engineering into something an auditor, a clinician, a quality function, or a regulator can examine and accept.

Defensibility has three faces. *Scientific defensibility* is evidence that the system is practically useful and works for its intended use and context. *Technical defensibility* is the set of safeguards built into the system to make outcomes predictable by preventing or managing failure, drift, bias, and attack. *Ethical defensibility* is the meaningful human integration and oversight that preserves accountability and judgement as automation increases. These are operationalized through three commitments that recur throughout our framing: evidence-based practice, governance by design, and human-led deployment.¹⁴

The three faces map cleanly onto the agent and software question, as summarized in Table 1.

Table 1. Three forms of defensibility applied to agents and AI-assisted software

DIMENSION	THE QUESTION IT ANSWERS	WHAT IT MEANS FOR AGENTS AND SOFTWARE
Scientific defensibility	Is the system practically useful and fit for its intended use and context?	Evidence that an agent performs its task, and that AI-assisted code meets its specification, proportionate to risk and sustained through monitoring.
Technical defensibility	Are outcomes predictable, with failure, drift, bias, and attack prevented or contained?	Guardrails, tool-gating, least-privilege access, reversibility, and audit trails engineered into the agent's execution path.
Ethical defensibility	Is human accountability and judgement preserved as autonomy increases?	Meaningful oversight, clear ownership and escalation, and literate users who know when to trust, verify, or override.

The organizing idea that makes this workable is proportionality. The same operating model applies to every system, but the bar is calibrated to the use. A commercial drafting assistant should attract promotional-compliance checks and drift monitoring, not a clinical validation package; an agent that touches a regulated release should attract the fuller weight of evidence and oversight. Proportionality is what prevents governance from becoming either theatre for low-risk uses or a rubber stamp for high-risk ones.

Finally, these expectations are best understood as a compliance floor rather than a ceiling.¹⁵ Building without them in mind does not avoid the cost, it defers it. The cost will need to be paid back later as retrofit, delay, and rework.^{16,17} The winning move is speed with defensibility built in rather than bolted on afterwards, which keeps development regulator-ready without slowing the cycle. The sections that follow describe the floor: how to tier risk, how to validate proportionately, and how to engineer the evidence of trust.

Proportionate risk tiering as a defining principle

Before deciding how much validation and control a system needs, we need to decide how much risk it carries. The foundational principle is that tiering is use-driven. The risk tier depends on the context and downstream use of the outputs, rather than the sophistication of the model. Drawing on model-risk-

management practice and the OECD dimensions for classifying AI systems, inherent risk generally rises with proximity to patient care, to product quality, and to regulated decisions, and falls for infrastructure-level analytics removed from clinical or regulatory impact.^{18,19} Tiers from foundational and administrative use through to patient-facing and safety-critical systems should give enough granularity to be meaningful while remaining easy to explain.

Risk areas	Evidentiary bar	Regulatory examples	Use cases
Clinical R&D	Heavy scientific validation; PCCPs for evolving models	FDA credibility framework, EMA reflection paper, GCP/GLP, TRIPOD	Trial design, site selection, patient recruitment, digital twins, safety, RWE
Commercial and medical	Promotional/PV/privacy compliance; HITL for MLR	FTC/EMA promotional, HIPAA/GDPR, PhRMA code	HCP engagement, DTC advertising, market access, med affairs, KOL
Manufacturing, supply chain, quality	Life-cycle maintenance within pharmaceutical quality system; continuous monitoring	GxP, ICH Q8/Q9/Q10, Annex 11, GAMP 5	Process control, batch release, deviation prediction, supply resilience
Enabling functions	Data protection, IP, privileged access; light governance	ISO/IEC 42001/5, Internal AI policy, Privacy law	Literature summarization, contract analytics, HR, finance, back-office

A single defensible AI operating model, applied *proportionately*, serves all four areas. Agent autonomy and execution will increase scrutiny.

A tiering framework also separates inherent risk, the baseline before controls, from residual risk, what remains after mitigation. This distinction, common to the EU AI Act, the NIST AI Risk Management Framework, and ISO/IEC 42001 AI management system, ensures that governance, validation, and oversight are proportionate to a system’s true potential for harm, and that the effect of controls is documented rather than assumed.^{4,10,20}

Autonomy is the amplifier that agents introduce

Model exposure, the class of model and its opacity and adaptivity, modulates but doesn’t define the tier. A useful overlay is to tier first by use and then adjust for exposure. A deterministic rules engine changes little, whereas a generative or foundation model in a

patient-facing context warrants escalation because it can amplify the risk, and therefore the controls and expected validation for the system.

This is where agents demand particular care. An agent is typically built on a generative model and is, by exposure, at the top of the scale, yet its tier is still driven by what it’s allowed to do. The decisive variable is autonomy, because autonomy moves a system from advice to execution.²¹ Industry evidence points the same way. In the 2026 revision of the most widely used catalogue of language-model application risks, excessive agency rose to become one of the top three concerns, the largest upward movement in the list, under a methodology that for the first time weighted analysis of documented real-world incidents alongside practitioner consensus.²²

The framework for tiering can capture this through risk modifiers. Critical amplifiers, such as the execution of regulated decisions, autonomous clinical or quality actions, or operation within GxP environments, are red-line conditions that escalate a system toward the highest tier regardless of its base use, because they remove the human, procedural, or temporal buffers that would otherwise contain an error. Significant amplifiers, such as the use of identifiable data, autonomous execution without human sign-off, or integration with electronic health records and other systems of record, can raise the tier by one level to reflect increased exposure while still allowing structured human oversight to manage harm. Factors such as third-party model use or cross-border data flows, however, may not change the tier. Instead, they can strengthen the controls applied within it.

A short illustration makes the logic concrete. An off-the-shelf assistant used internally to prepare meetings, summarize documents, and track actions is built on a foundation model but operates on non-regulated data with advisory outputs, and sits at the lowest tier. The generative exposure alone doesn't necessarily escalate it. Wire that same generative capability into a workflow where it executes against a regulated system of record, and the autonomy amplifier applies because the system now behaves as a higher-tier deployment and attracts the corresponding validation and oversight. The capability didn't change. The use, and therefore the risk, did. Table 2 summarizes a tiering posture at the level of principles.

Table 2. Risk tiering principles at a glance (example only)

RISK BAND	TYPICAL USE AND PROXIMITY TO IMPACT	AUTONOMY AND EXPOSURE	VALIDATION AND OVERSIGHT POSTURE
Lower	Administrative, infrastructural, or exploratory; advisory outputs; removed from regulated decisions or patient impact.	Advisory only; human acts on suggestions; no execution on systems of record.	Lightweight documentation, version control, and periodic review; over-engineering is discouraged.
Moderate	Translational, pre-clinical, or non-critical operational optimisation with human approval.	Constrained tool use; human review before action; bounded write access.	Systematic technical evaluation, provenance, bias screening, and defined monitoring.
Higher	Clinical, regulatory, or high-impact commercial decisions; patient-facing or product-quality actions.	Autonomous or semi-autonomous execution; integration with regulated workflows escalates the tier.	Workflow- and real-world evaluation, independent challenge, continuous monitoring, and executive sign-off.

Scientific validation as proportionate evidence

Validation is a body of evidence assembled across the lifecycle, scaled to the tier. Across the international frameworks that matter here (such as the FDA and EMA guidance, TRIPOD+AI and TRIPOD-LLM, GAMP AI, NIST, and ISO/IEC 42001 AI management system and ISO/IEC 42005 AI system impact assessment), a consistent set of principles emerges:^{1,2, 4,5,6,7,8,20} risk-based proportionality, lifecycle integration, multi-dimensional assessment, an external-validation imperative that grows with tier, fairness and equity, transparency and explainability, continuous monitoring, and stakeholder engagement. In practice, evidence can be organized across the dimensions of data, model, analytical, clinical, and operational, whose depth scales with risk rather than applying uniformly to every system.

Validating an agent, however, is different than validating a model. A model has a bounded input and output that can be characterized on a held-out dataset. An agent has emergent behavior, a dynamic execution path, tool use, and the capacity for errors to propagate across steps and compound. Technical evaluation of the underlying model, its accuracy, calibration, and robustness, remains necessary but is no longer sufficient. Evidence must extend to the workflow, evaluating the agent in the actual task, exercising each key step, assessing how errors propagate, and testing the system together with the human who supervises it.

A staged view of the principles in action

A defensible validation program moves through stages with explicit decision gates, so that progression depends on evidence rather than momentum. It begins with use-case characterization, defining the intended use and context, the decisions the system influences, the boundary between AI autonomy and human responsibility, and the inherent risk before controls. It proceeds to technical evaluation of the model and its data, then to workflow and task-level evaluation of the assembled system, then to a real-world pilot appropriate to risk, and finally to continuous monitoring and lifecycle governance. Each stage can conclude with a reviewer's judgement of whether the evidence is sufficient and whether residual risk is acceptable.

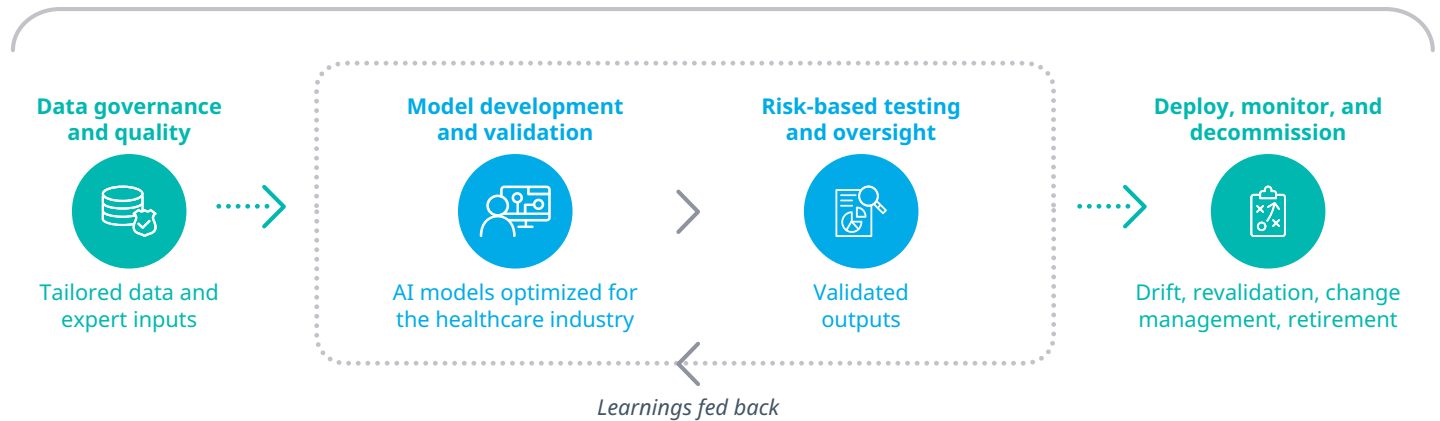
We describe these stages as principles precisely because the value on this subject lies in the reasoning rather than the artefacts. The detailed, tier-specific evidence expectations that operationalize these stages exist as a separate instrument for teams to apply. Reproducing them here would shift a discussion of principles into a procedures manual that needs organizational context that could otherwise obscure the ideas that leaders most need to internalize. The essential idea is to match the rigor of the evidence to the consequence of the action, and gather that evidence across the lifecycle rather than at a single moment.

Approaches to trust using evidence as by-product

Principles become defensible only when they are engineered into the system. Governance that lives only in policy will struggle at scaling to software that's written and run by agents. Governance has to be embedded

so that controls are enforced by design and produce evidence as a by-product of operation. Four engineering approaches carry most of the weight, and together they answer the practical question of how a customer can trust both the software an agent produces and the actions an agent takes.

Human oversight and expertise required at all stages



Audit logs and traceability

An agent's trustworthiness is only as strong as its audit trail. We can only govern what we can observe, and we can only attribute what we log. Traditional application logs, designed for deterministic software, record that an interface returned a result. But of the reasoning that led an agent to call a particular tool, or the chain that links a user's original intent to the agent's sub-tasks? Defensible agents therefore need intent-based, identity-linked telemetry that treats the agent as a first-class actor and captures what it did, why it did it, which tools and data it touched, and which model version, prompt, and configuration were in force.^{21,23} In GxP contexts this is also the substrate of data integrity, supporting traceability from source to output and the reconstruction that's demanded.⁸ Well-instrumented logs turn oversight from an interruption into a continuous, reviewable record.

User interfaces as trust surfaces

Oversight can happen at the interface where a person meets the system's output. The design of that surface largely determines whether human review is meaningful or a formality. Interfaces that communicate uncertainty, surface provenance and citations, and make it easy to inspect what an agent is about to do allow reviewers to calibrate their trust rather than defer to the machine.²¹

But an interface can only support oversight if the workflow leaves room for it. Even a well-informed reviewer can't evaluate an output delivered in a compressed decision window or buried among competing alerts, and defaults that make an override slower than acceptance quietly push behavior toward passive confirmation. Preserving this cognitive space, by managing alert burden and review latency, and by ensuring that questioning the system is at least as easy as accepting it, is as much a determinant of meaningful oversight as the accuracy of the underlying model.^{21,24}

Patterns such as approval-before-action for consequential steps, structured review with a recorded rationale, and overrides that are easy to exercise and automatically captured are how accountability is preserved in practice. Good interface design is also the most direct defense against automation bias, the tendency to accept a confident output without scrutiny, and it produces its own evidence in the form of the record of what was shown, reviewed, edited, and overridden.^{24,25}

These patterns are beginning to consolidate and are familiar ground in healthcare. Usability engineering is a regulated discipline for high-risk systems, in which interface design is specified and evaluated prospectively during development, documented for retrospective audit, and revised as post-market experience accumulates.^{26,27} What is new is extending that discipline to AI, where regulatory expectations already call for interfaces that make oversight possible, and where the corresponding design standards are still being distilled from regulation and industry practice.^{25,28}

Control by design through guardrails, policy-as-code, and tool-gating

For agents, guardrails must govern behavior, not just words. Content filters that moderate text are insufficient when the risk is an action, a tool call, a write to a system of record, or an authorization. Defensible design constrains the execution path through least-privilege access so an agent can reach only what its task requires, tool-gating and policy-as-code so that prohibited actions are prevented rather than detected after the fact, explicit action boundaries, reversibility and rollback for consequential steps, and a reliable stop or kill switch. These controls implement technical defensibility directly and, because they operate in the execution path, they also generate the signals that monitoring and audit depend upon.^{21,23,29}

Containment matters as much as prevention, because in multi-agent settings a single fault can propagate across agents and workflows faster than stepwise human checks can intervene. Controls such as quotas, breakers between planning and execution, and separation of the planning and execution paths can keep a local error from becoming a systemic one.²¹

Emerging open standards for runtime agent control now describe these enforcement points in portable form, though the work is early. Emerging work on runtime agent control describes declarative policies enforced through middleware hooks at each decision point, where a pending action is evaluated and allowed, denied, or modified before it reaches a production system, with structured traces emitted through conventional observability pipelines. The distinction it draws is the one that matters here, between observing what an agent did after the fact and intercepting what it's about to do, though the specification remains an early public preview as these concepts mature.³⁰

Using agents to validate agents

It's tempting to answer every concern about autonomy with the phrase "human in the loop," but presence in the loop does not by itself make oversight meaningful. A reviewer who lacks the knowledge, time, authority, or practical means to intervene supplies the appearance of control rather than its substance.²⁴ Oversight is a property of the system and the organization around it, not a virtue expected of an individual at the point of use. Designing agents to validate agents should be a way to concentrate human judgement where it is decisive, and the discipline described below is defensible only if these conditions hold.

Nor is oversight binary. Between reviewing every action and ceding the field lie graded positions: a human approving each output before it takes effect (human in the loop), a human monitoring an otherwise autonomous system and intervening by exception (human on the loop), and a human governing the parameters within which a system may operate at all (human in command). These modes are long established and are not mutually exclusive, and guidance is explicit that oversight should be commensurate with the risk, the level of autonomy, and the context of use.^{25,31} The same proportionality that governs tiering. Establishing when a system has earned movement between these positions, and what harness of controls, monitoring, and accountability must hold for it to do so, is a question we take up separately.

The volume of software that agents can produce would overwhelm manual validation, which is why validation itself must be partly automated, using agents to accelerate the work while humans remain accountable for the judgement. Agents can generate and run regression tests, assemble evidence, map controls to regulatory requirements, and conduct adversarial probing such as prompt-injection and robustness testing, at a scale and speed no manual process can match.^{21,32} The discipline that keeps this defensible is human review at the points that matter. People should set the acceptance criteria, adjudicate edge cases, and accept residual risk, while agents do the labor of producing and organizing evidence. This is how validation keeps pace with agentic development rather than becoming its bottleneck, and it's a capability, not merely a control.

PUTTING THE PRINCIPLES TO WORK

Adopted in stages, each stage produces evidence the the next one builds on.

STAGE	WHAT TO PUT IN PLACE	EVIDENCE IT PRODUCES
Foundations	An inventory of AI systems and agents, a use-driven risk tier for each, clear ownership and escalation, and role-appropriate literacy.	A defensible basis for deciding how much validation and oversight each system needs.
Operationalization	Validation proportionate to tier, oversight patterns matched to consequence, monitoring and drift detection, incident response and rollback.	Proof that systems work as intended at launch and remain within bounds.
Advanced	Policy-as-code and tool-gating in the execution path, automated checks, audit trails that capture what agents did and why.	Controls that prevent rather than detect, with evidence produced as a by-product of operation.
Ongoing	Override and discordance review, revalidation triggers, regulatory updates, lessons learned.	Assurance that is maintained rather than certified once.

We've applied these principles in our own work and with clients. In one engagement, an organisation under board scrutiny over its use of data and AI adopted use-case analysis and risk tiering within its existing model risk management. This gave it a clear decision framework for which systems needed which level of validation and oversight, without building a new governance structure. In another, we used agents to accelerate the mapping of AI controls to regulatory and best-practice requirements, while domain experts set acceptance criteria and adjudicated the results. That is the division of labour this brief describes, where agents produce and organise evidence, and people stay accountable for the judgement.

Lifecycle governance and continuous assurance

Because agents adapt and their environments change, trust must be maintained rather than certified once. This is where ethical defensibility is sustained rather than asserted, because accountability, authority, and understanding either persist through the life of a system or quietly erode as its use becomes routine.²¹

Continuous assurance rests on a few durable commitments:^{4,5,20}

- Clear ownership and accountability for every deployed system
- Stage gates that require evidence and explicit residual-risk acceptance before progression
- Monitoring for performance and drift with cadence proportionate to tier
- Defined revalidation triggers
- Controlled change management for models, prompts, tools, and configurations
- A plan for retiring a system when its risks can no longer be managed

Oversight requires decisional authority, which is an organizational commitment. People must be supported to question, override, or escalate an agent's output without that divergence being treated as deviance, since justified disagreement that's expected, documented, and protected is what separates real oversight from a symbolic sign-off.²⁴ Override and discordance should likewise be read as safety signals rather than compliance failures.

An oversight view that tracks override frequency, review latency, agreement and disagreement between agent and human, subgroup performance drift, and near misses turns the audit trail into an early-warning system, provided these indicators trigger interpretation and follow-up rather than serving as standalone scorecards.

What they are ultimately watching for is both whether a system is degrading and also whether anyone is being harmed by it, and the subgroup view matters precisely because harm is rarely distributed evenly.²⁴

Meaningful human oversight is only possible where the humans exercising it understand the systems they supervise, a point made unavoidable by Article 4 of the EU AI Act.^{10,11} This understanding is specific, not general. The people relied upon to oversee a system need to know its intended use, the populations and conditions in which it was validated, how its performance varies across subgroups, and where it's most likely to fail, otherwise a reviewer is left to choose between blind trust and blanket rejection. Transparency artefacts such as model cards only help when paired with training, onboarding, and support rather than offered on their own.

Literacy is more than a training checkbox. Literacy is an oversight control and a market-access consideration, as Gartner observes.^{12,13} The obligation is proportionate, calibrated to a person's role, the risk of the system, and the people affected, which mirrors the tiering logic of everything above. Building AI literacy into how teams work is therefore part of the operating model, instead of an adjacent HR program, and the record of it is part of the evidence that a system can be trusted.

The heavier instrumentation belongs to higher tiers, where real-time monitoring, automated rollback, and post-market-style surveillance are appropriate. Lower tiers are served by version control and periodic review. What matters is that oversight scales with consequence, so that judgement remains with people as autonomy grows.



Conclusion

Software development has changed, and healthcare and life sciences are being asked to trust software that agents increasingly write, test, and run. As systems move from advising to acting, trust has to be produced. Trust is produced as scientific evidence proportionate to risk, as technical controls engineered into the execution path, and as the literate human judgement that keeps accountability with people. A defensible operating model brings these together and applies them proportionately, so that the bar fits the use rather than defaulting to the most conservative setting for everyone.

The organizations that pull ahead will not be the ones with the most autonomous agents. The organizations that pull ahead will be the ones that can show their work, whose agents carry an audit trail, whose interfaces make oversight real, whose controls prevent rather than merely detect, and whose people understand what they are supervising. That's what it means to build agents and software that earn adoption, withstand scrutiny, and remain accountable across their lifecycle. It is, in a phrase, defensibility built in rather than bolted on.

References

1. European Medicines Agency. Reflection paper on the use of artificial intelligence in the medicinal product lifecycle [Internet]. EMA; 2024 Sep. <https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle-scientific-guideline>.
2. U.S. Food and Drug Administration. Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products (draft guidance) [Internet]. FDA; 2025 Jan. <https://www.fda.gov/media/184830/download>.
3. U.S. Food and Drug Administration, European Medicines Agency. Guiding principles: good machine learning /AI practice for drug development [Internet]. 2026 Jan. <https://www.fda.gov/about-fda/artificial-intelligence-drug-development/guiding-principles-good-ai-practice-drug-development>.
4. International Organization for Standardization. ISO/IEC 42001:2023 — AI management system [Internet]. ISO; 2023 Dec. <https://www.iso.org/standard/42001>.
5. International Organization for Standardization. ISO/IEC 42005:2025 — AI system impact assessment [Internet]. ISO; 2025 May. <https://www.iso.org/standard/42005>.
6. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. <https://www.bmj.com/content/385/bmj-2023-078378>.
7. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60–69. <https://www.nature.com/articles/s41591-024-03425-5>.
8. International Society for Pharmaceutical Engineering. GAMP guide: artificial intelligence [Internet]. ISPE; 2025 Jul. <https://guidance-docs.ispe.org/doi/book/10.1002/9781946964854>.
9. European Federation of Pharmaceutical Industries and Associations. AI across the medicines lifecycle: insights and considerations for policy [Internet]. EFPIA; 2025 Nov. <https://efpia.eu/news-events/the-efpia-view/efpia-news/ai-across-the-medicines-lifecycle-insights-from-preliminary-case-studies-and-considerations-for-policy/>.
10. European Parliament. EU AI Act: first regulation on artificial intelligence [Internet]. 2024 Jun. <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>.
11. European Commission. AI literacy – questions and answers (Article 4, AI Act) [Internet]. Shaping Europe’s digital future; 2025. <https://digital-strategy.ec.europa.eu/en/faqs/ai-literacy-questions-answers>.
12. Zimmerhansl S, Roussel A. First take: the EU AI Act’s Article 4 just turned AI literacy into a business risk [Internet]. Gartner; 2026 Aug 3. <https://www.gartner.com/en/documents/8208729>.

13. Gartner. AI literacy is no longer optional: the business impact of EU AI Act Article 4 [Internet, webinar]. 2026. <https://www.gartner.com/en/webinar/960615/1951664-ai-literacy-is-no-longer-optional-the-business-impact-of-eu-ai-act-article-4>.
14. Arbuckle L. A blueprint for defensible AI [Internet]. IQVIA; 2024 Oct. <https://www.iqvia.com/blogs/2024/10/a-blueprint-for-defensible-ai>.
15. OWASP GenAI Security Project. GenAI security industry framework crosswalk [Internet]. 2026 Sep 1. <https://genai.owasp.org/resource/genai-security-industry-framework-crosswalk/>.
16. Boston Consulting Group. The widening AI value gap: build for the future 2025 [Internet]. BCG; 2025 Sep. <https://www.bcg.com/publications/2025/are-you-generating-value-from-ai-the-widening-gap>.
17. Challapally A, Pease C, Raskar R, Chari P. The GenAI divide: state of AI in business 2025 [Internet]. MIT Project NANDA; 2025 Jul. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf.
18. Kiritz N, Ravitz M, Levonian M. Model risk tiering: an exploration of industry practices and principles. J Risk Model Validation. 2019. <https://www.risk.net/journal-of-risk-model-validation/6710566/model-risk-tiering-an-exploration-of-industry-practices-and-principles>.
19. Organisation for Economic Co-operation and Development. OECD framework for the classification of AI systems. OECD Digital Economy Papers No. 323. Paris: OECD Publishing; 2022. <https://doi.org/10.1787/cb6d9eca-en>.
20. Tabassi E. Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg (MD): NIST; 2023 Jan. <https://doi.org/10.6028/NIST.AI.100-1>.
21. OWASP GenAI Security Project. OWASP Top 10 for agentic applications 2026 [Internet]. 2025 Dec. <https://genai.owasp.org/>.
22. OWASP GenAI Security Project. Top 10 for LLM applications 2026 [Internet]. 2026 Aug 3. <https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/>.
23. OWASP GenAI Security Project. Securing agentic applications guide v1.0 [Internet]. 2025 Jul 28. <https://genai.owasp.org/resource/securing-agentic-applications-guide-1-0/>.
24. Van de Sande D, Economou Zavlanos N, van Genderen ME. Meaningful oversight of medical AI beyond human in the loop. npj Digit Med. 2026;9:569. <https://doi.org/10.1038/s41746-026-02971-1>.
25. European Parliament and Council. Regulation (EU) 2024/1689, Article 14: human oversight [Internet]. <https://artificialintelligenceact.eu/article/14/>.
26. International Electrotechnical Commission. IEC 62366-1:2015 — Medical devices, Part 1: application of usability engineering to medical devices. Geneva: IEC; 2015. <https://www.iso.org/standard/63179.html>.

27. U.S. Food and Drug Administration. Applying human factors and usability engineering to medical devices: guidance for industry and FDA staff [Internet]. FDA; 2016 Feb. <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/applying-human-factors-and-usability-engineering-medical-devices>.
28. European Commission. Standardisation of the AI Act [Internet]. Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/policies/ai-act-standardisation>.
29. National Institute of Standards and Technology. Cybersecurity framework 2.0. NIST CSWP 29; 2024. <https://doi.org/10.6028/NIST.CSWP.29>.
30. OWASP GenAI Security Project. Agent Control Standard (ACS) v0.1 [Internet]. 2026 Sep 1. <https://genai.owasp.org/resource/agent-control-standard-acs/>.
31. High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI [Internet]. European Commission; 2019 Apr 8. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.
32. MITRE. ATLAS: adversarial threat landscape for artificial-intelligence systems [Internet]. 2025. <https://atlas.mitre.org>.

About the authors



LUK ARBUCKLE

Global AI Practice Leader in Defensible Data & AI,
AI & Technology Solutions, IQVIA

Luk is an AI and data security strategist with 25+ years in applied AI science and over a decade in privacy engineering, spanning industry, regulatory, and standards bodies. He designed IQVIA's enterprise AI risk tiering, scientific validation tiering, and risk mitigation frameworks, mapped to NIST, ISO/IEC, and international regulatory guidance, and operationalized as proportionate, auditable controls across the model lifecycle. Formerly Director of Technology Analysis at Canada's federal data protection authority, he has co-authored two O'Reilly books, holds multiple patents, and has 10 years' experience in international standards development. Luk provides strategic leadership to clients and to IQVIA on the methods to get the most value from data with defensible AI.



GARETH DABBS

Vice President of Enterprise Platforms,
AI & Technology Solutions, IQVIA

Gareth leads strategy formation, portfolio management, AI innovation, partnerships, and technology transformation initiatives across the AI & Technology Solutions (ATS) organization. Gareth focuses on applying AI, data, and emerging technologies to accelerate value creation for life sciences organizations across the clinical and commercial domains. Work focuses on trusted AI, enterprise knowledge systems, strategic partnerships, and the adoption of next-generation workflow, analytical and agentic capabilities across client organizations.

About IQVIA

IQVIA (NYSE:IQV) is a leading global provider of clinical research services, commercial insights and healthcare intelligence to the life sciences and healthcare industries. IQVIA's portfolio of solutions are powered by IQVIA Connected Intelligence™ to deliver actionable insights and accelerate innovations. With approximately 91,000 employees in over 100 countries, IQVIA is dedicated to accelerating the development and commercialization of innovative medical treatments to help improve patient outcomes and population health worldwide. Learn more at www.iqvia.com.



CONTACT US
defensibleAI@iqvia.com
iqvia.com/contact