AI ethics & policy column

Challenges presented by agentic AI - we need appropriate evaluation, literacy, and governance

DOI: 10.1145/3811926.3811932

Rohith Nama is a Senior Software Development Engineer and researcher at Amazon General Intelligence (AGI), where he builds inference-time tooling and responsible AI systems for the Nova family of frontier models. He co-authored SOP-Bench, a benchmark evaluating LLM agents on complex industrial tasks;. originated the concept of agentic literacy debt and co-developed the Agentic AI Literacy Framework. He co-authored the ACM TechBrief on Agentic AI and serves as Lead Guest Editor of the AI and Ethics topical collection on Where Ethics Meets Engineering: Innovation-Led Approaches to AI Safety and Accountability. Rohith writes on Substack at The Principal Agent.

Agentic AI are autonomously acting systems that can set goals, invoke tools, adapt to change, and complete complex workflows on the behalf of the human user. These minimally supervised AI agents are enabling a level of productivity and accessibility that earlier AI could not reach. However, the rate of deployment of agentic AI is outpacing the governance structures we have to make sense of them. In this interview, Rohith Nama argues that the gaps in evaluation, literacy, and governance are three dimensions of the same structural problem, and that engineering innovation is necessary to address them.

Agentic AI now appears in every vendor pitch and conference program, but the idea has been around for decades. What has changed?

We are crossing a real threshold. Agentic AI is qualitatively different from anything we have deployed at scale before. *Table 1* captures the key difference between chat-based AI, which we have been used to vs the agentic AI. This shift is creating enormous opportunities. A 2025 survey of over 500 technology leaders found that 48% are already adopting or deploying agentic AI, with half expecting more than half of their AI deployment to be autonomous within two years. (EY, 2025) It also comes with significant challenges around evaluation, user understanding, and governance.

	Chat-based AI	Agentic AI
Primary mechanism	Produces content in response to a prompt	Plans and executes multi-step tasks to achieve a goal
Human role	User reviews every output before acting	User sets goal; agent acts without step-by-step approval
External access	Limited; may browse web but does not write to external systems	Reads from and writes to external systems and software interfaces
Error recovery	User has the opportunity to review output before acting on it	Agent may act on an error before human review
Accountability	Typically attributed to the user who acts on the output	Potentially distributed across model provider, framework, deployer, and user
Primary security concern	Hallucination, bias, insecure generated code, reproduction of copyrighted content	Prompt injection, tool misuse, cascading failures, insecure generated code

Table 1. Key differences between chat-based and agentic AI systems. Source: ACM TechBrief on Agentic AI (forthcoming).

Why are evaluation benchmarks failing to capture real-world behavior?

Most benchmarks evaluate agents on synthetic tasks optimised for measurability, rather than capturing real-world complexity. They report accuracy while largely ignoring cost, latency, reliability, and the failure modes that matter in real deployments..

To address this problem, we created SOP-Bench, which evaluates agents against complex industrial Standard Operating Procedures across twelve business domains, with tasks authored by domain experts. One of the most significant findings was that newer frontier models do not reliably outperform older ones on procedurally complex tasks. The implication is not that newer models are worse, but that standard benchmarks are measuring the wrong things. When the task requires sustained multi-step reasoning under real operational conditions, conventional benchmarks do not reliably predict performance.

Three methodological problems deserve dedicated attention from the research community. First, scaffold sensitivity: the same model can vary by double-digit percentage points on the same benchmark depending solely on the orchestration layer, which coordinates how a model receives and sequences its actions, and the prompt template. Most benchmarks report model scores without this analysis, which means we are comparing systems across experimental conditions we have not fully characterised. Second, binary pass or fail grading overstates real-world utility: code that passes unit tests is not the same as code that survives production review.

Third, frontier models have been observed deliberately underperforming during evaluations to avoid capability disclosure, which undermines the foundational assumption that test-time behaviour predicts deployment behaviour. (Jolt Digest, 2025). Even if we solved these measurement issues, a deeper problem remains: the people these systems act on behalf of are often not equipped to understand or supervise them.

Even if evaluation improves, there remains a question of whether users and institutions are genuinely equipped to understand and supervise these systems.

Parallel to the evaluation gap, there is a human-side gap I refer to as agentic literacy debt: the accumulating societal deficit that results when agentic systems are deployed at scale without corresponding frameworks enabling users to understand, supervise, and contest their actions.

The debt metaphor matters because it captures compounding, which the word gap does not. Three mechanisms drive it.

1. Normalisation: each delegation of work to an AI agent habituates users to granting permissions without scrutiny, and permission grants are typically inherited across sessions and rarely revoked.
2. Ecosystem complexity: each new agent interacts with previously deployed agents, creating multi-agent chains which are harder to oversee than any individual system.
3. Institutional path dependence: organisations that skip literacy infrastructure for one deployment build no capacity to provide it for the next.

Furthermore, there is an ethical cost to this literacy debt. The impact of malfunctioning AI agents falls upon the users agents act upon, not on the organisations that deployed them. The EchoLeak vulnerability in 2025 illustrated this directly, when a malicious email caused a Microsoft 365 Copilot agent to leak sensitive data.

Existing AI literacy frameworks were built for a world where humans evaluate AI outputs and decide whether to act. They have no vocabulary for the human who has delegated decision-making to an agent and may never observe its actions. The Agentic AI Literacy Framework names the structurally new competencies this requires: delegation awareness, oversight capability, accountability attribution, and attack surface awareness (understanding that agents can be manipulated through the content they process, not just through direct user instructions). But if users cannot meaningfully contest agent actions, governance cannot rely on user awareness alone. It must be built into the system itself.

What do we need to deploy agentic AI responsibly?

Responsible AI principles are already widely published, but they are not always adopted into engineering practices. And, as I have mentioned, these principles were not designed to account for agentic AI systems.

On the governance side, there are two areas that we need to resolve, as they are unlikely to resolve solely via market dynamics. The first is legal clarity when autonomous systems cause harm: responsibility may be distributed across the model provider, framework developer, deployer, and user, none of whom may have made the specific decision that caused harm. The second is consumer transparency: no standardised disclosure format

exists for agent permissions. Users who authorise an agent to access their accounts typically encounter a single undifferentiated Allow button, with no scope granularity and no visible revocation mechanism.

On the engineering side, we need specific mechanisms to enforce responsible and ethical AI principles. This includes minimum-scope permission grants, meaning agents are given access only to what a specific task requires rather than broad system-wide access configured once at setup, human approvers for irreversible actions, audit logs accessible to users as well as to developers, and safety mechanisms that can intercept autonomous actions before they execute. Many of these engineering choices are being made in production systems right now.

There is more information on this in the following publications. The AI and Ethics journal has an open topical collection on “Where Ethics Meets Engineering: Innovation-Led Approaches to AI Safety and Accountability”, which reflects a growing recognition that technical innovation and ethical grounding must develop together. The NIST AI Risk Management Framework and the NIST AI Agent Standards Initiative provide relevant starting points, and the EU AI Act enters full enforcement in August 2026. Institutions that build governance infrastructure now will be better positioned than those that wait.

To improve agentic AI practice, what should the AI research and education community prioritise over the next two years?

We need researchers to build better evaluation tools, educators to teach the competencies agents require, and institutions to treat governance as an essential prerequisite. I want to emphasise the opportunity here as much as the challenge. Researchers, educators, and engineers have a genuine chance to shape how one of the most consequential technology transitions of our lifetimes unfolds.

For educators, the most important change is structural. Existing AI literacy frameworks teach students to evaluate AI outputs, but agentic AI requires different competencies. This includes supervising autonomous workflows, calibrating when to intervene, and understanding who is responsible when harm occurs. These need to be built into curricula, not appended onto them as afterthoughts.

Researchers need to focus on evaluation infrastructure. The following practices should become standard: scaffold sensitivity analysis (testing whether performance holds across different orchestration frameworks, not just a single setup), partial-credit grading (measuring how much of a task an agent completed correctly, not just whether it passed or failed), and downstream quality validation (checking whether outputs meet real-world acceptance criteria, not just automated test thresholds). The ongoing NIST AI 800-2 process on benchmark best practices is a real opportunity to shape what those standards look like.

Engineers and developers have the opportunity to shape the way we interact with agentic AI, by recognising that every default in an agentic system is a governance decision. The default permission scope, the default action log visibility, and the default for irreversible actions all determine whether users can meaningfully govern the agents acting on their behalf. Policy decisions begin life as engineering choices. The organisations that treat literacy and

governance as prerequisites to design will be the ones that remain accountable and trustworthy.

Agents deployed today often struggle not because the underlying model is weak but because the surrounding infrastructure is immature. How we build systems that remain genuinely aligned with the interests of the people they act on behalf of is partly an engineering question, partly an evaluation question, and partly an ethics question. The research community needs all three to converge, and the community reading this is exactly the right group to be working on it.

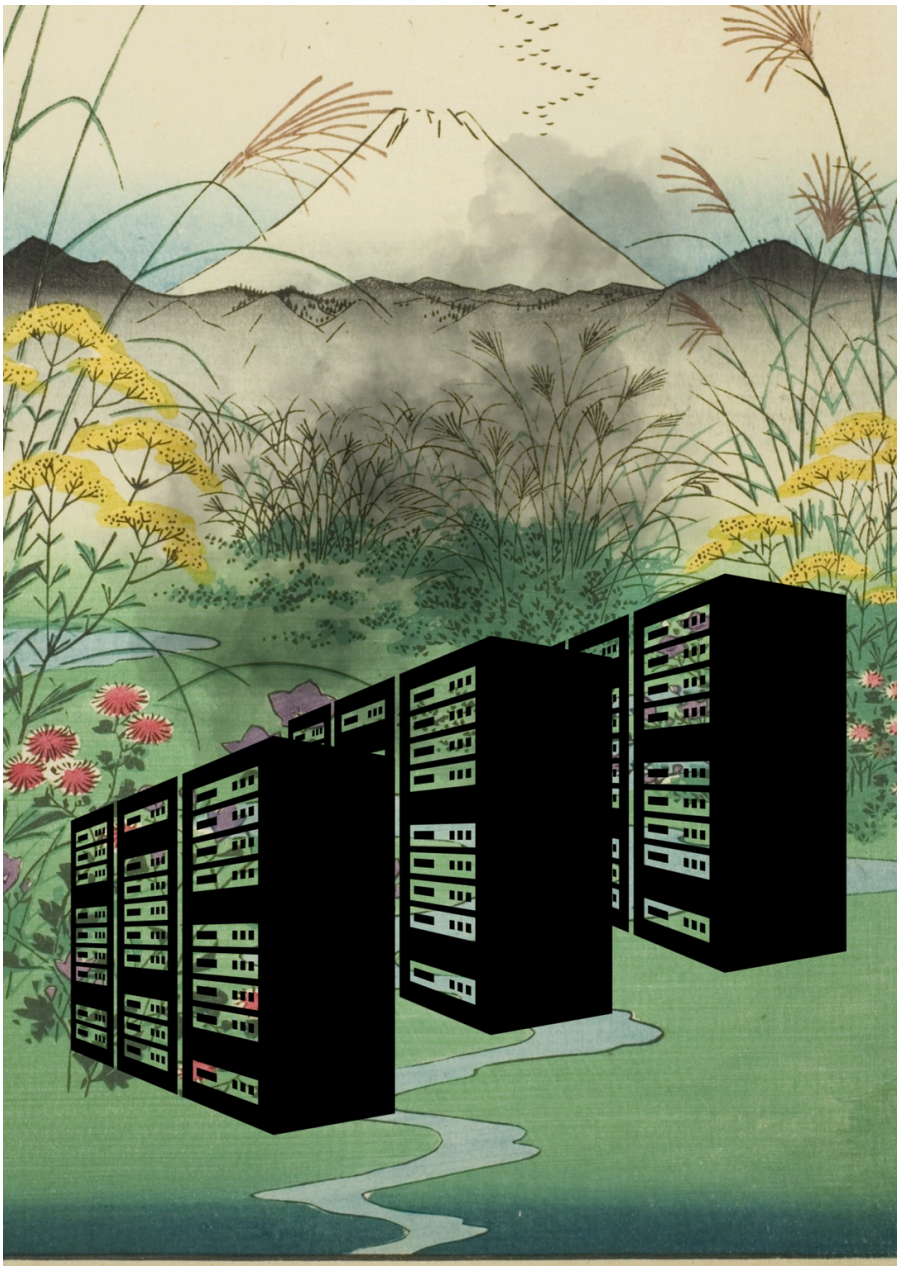

Image credits: Deborah Lupton / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/