Why Agentic AI Was Always Going to Fail

The Inevitable Failure of Regulate AI and Stop AI Is Why We Are Converging on Augmented Intelligence

Basil C. Puglisi, MPA
A Human-AI Collaboration

basilpuglisi.com | June 2026
#AIassisted using HAIA Ecosystem

Abstract

The agentic AI era promised to replace humans. Systems would plan, choose tools, and act on their own, with no named human holding binding authority over the output. That promise was always going to fail, and it has failed on two fronts. The technology has not delivered reliable human replacement: the best autonomous agents complete roughly a quarter of common workplace tasks, multi-step accuracy collapses by nearly half, and over 40 percent of agentic AI projects are forecast for cancellation by 2027. At the same time, the public is rejecting the premise itself. Supermajorities in 2026 polling oppose AI making consequential decisions without a human accountable for the result, and that opposition holds across enterprise policy and new law on both sides of the Atlantic.

Agents are not useless. Agents as unchecked replacements for human judgment, autonomous entities running free of named human authority, have proven neither reliable nor acceptable. The field routinely collapses two variables into one: the control of a process and the location of accountability. A single test, whether a named human holds binding authority over the output, sorts any system into Responsible AI (automation and agents operating without a named human bound to the output) or AI Governance without reference to how autonomous it is. Agents governed by a named human remain a valid and productive technology. Agents released as human replacements are the era that was always going to fail.

The hypothesis

Agentic AI was always going to fail as a replacement paradigm.

The subject is specific. The Agentic AI Era is the project to replace humans with autonomous systems: minimal human supervision, the human removed from the consequential decision, and no named human holding binding authority over the output. The industry's dominant definitions point in this direction. IBM defines agentic AI as systems that reason, plan, use tools, and act toward a goal with minimal or no human supervision. EY, Gartner, and the major enterprise vendors converge on the same feature: the absent human.

That project has failed twice over, because the technology has not delivered what it promised. Autonomous agents cannot reliably replace the humans they were meant to displace, and the production evidence is not ambiguous. At the same time, the public, markets, and lawmakers are rejecting the replacement premise even where the automation partially works, because accountability when the task goes wrong matters more to the public than whether the machine can do the task at all.

Three scope conditions discipline the argument and must be present every time the failure is asserted.

First, agents are not the problem. Agentic capability, the technical capacity to plan, select tools, act through external systems, and pursue goals across multiple steps, is real and will improve. Agents under named human authority are a productive technology. The failure is the specific project of replacing humans with agents that run free, create outputs without oversight, and operate as autonomous entities unchecked by a person who answers for the result.

Second, the technology failure is about replacement, not about capability in general. Factory-mode automation succeeds in many bounded applications and will continue to. What the production record shows is that agents cannot yet reliably take over the consequential human decisions they were sold as replacing. A system that completes 24 percent of workplace tasks is a useful tool under supervision. It is not a replacement for the human it was marketed as displacing.

Third, Responsible AI, meaning automation and agents operating without a named human bound to the output, has a legitimate place. Factory mode is a valid organizational choice where stakes allow and outputs are reversible. Responsible AI did not fail. The replacement paradigm has failed, both technically and publicly, and selling factory mode as governance is the mislabeling that fuels the backlash.

The two axes

Most confusion in the field comes from collapsing two independent variables into one. The market version of that confusion is agent-washing: the practice of selling autonomy as if it were governance. The framework rests on separating them.

Axis A is process control: who directs the steps. This axis runs from scripted (a human authored the path) to autonomous (the model directs its own path). The industry calls this the difference between automation and agentic. It is a real engineering distinction that matters for reliability and debugging, and it affects cost. It does not by itself determine accountability.

Axis B is accountability: who answers for the output. This axis runs from no named human to a named human with binding authority through four channels: moral (answerability to the ethical standards each culture holds), professional (the named human's industry, licensing body, or appointing authority can remove their ability to practice), civil, and criminal. This is the line between Responsible AI (automation without a named human bound to the output) and AI Governance (a named human holds binding authority over the result), and it is the only axis that determines governance.

A recurring confusion in the field comes from treating these as one variable. When a vendor says a system is "agentic," the listener hears both more autonomous and less governed at once, because the industry's own definition ties the two together by specifying minimal human supervision. The framework separates them and holds them apart.

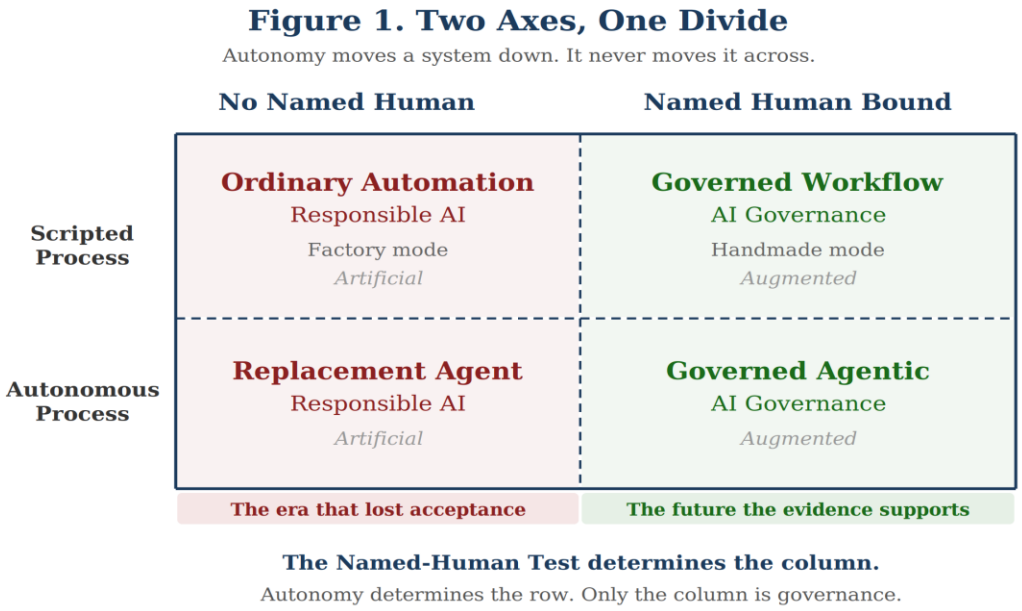

Figure 1. Two Axes, One Divide

Read the columns of the figure, not the rows. Both left cells are Responsible AI regardless of whether the process is scripted or autonomous, and both right cells are AI Governance regardless of whether the process is scripted or autonomous. Autonomy moves a system down a column, increasing technical independence, but it never moves a system across to governance because only a named human does that.

The left side of the figure is legitimate. Factory mode is a valid choice where stakes allow and outputs are reversible. The left column carries no condemnation. What failed is the specific promise that the left column could replace the right one, and the mislabeling of factory mode as if it were governance. The failure belongs there, not in agentic capability itself.

The Named-Human Test

The Named-Human Test sorts any system into Responsible AI or AI Governance with a single question, and the evidence that follows tests whether the distinction holds across markets and polling and through enacted law.

The test is a single question:

> *Is there a named human who holds binding authority to approve, modify, or halt the output, and whose accountability survives an audit through at least one of the four channels (moral, professional, civil, criminal)?*

If no, the system is Responsible AI (automation without a named human bound to the output) regardless of how autonomous it is. If yes, the system is AI Governance regardless of how autonomous it is. In plain terms: who can stop it, and who answers for it?

Figure 2. The Named-Human Test

A single classifier for any system, vendor claim, or build

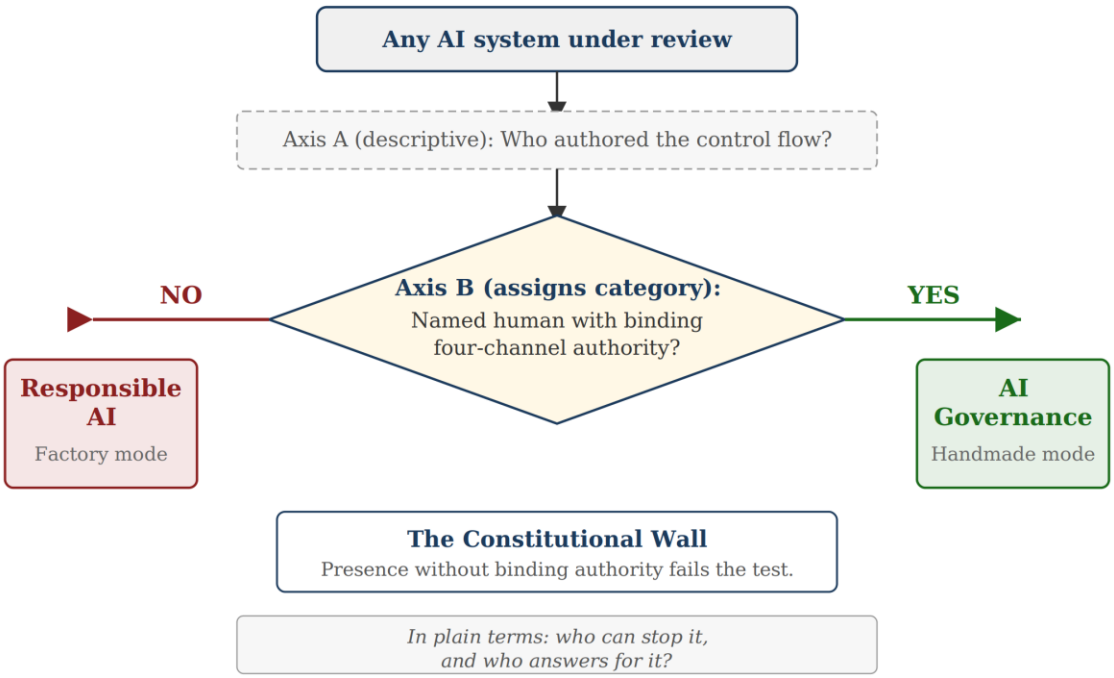

Figure 2. The Named-Human Test

The test includes a high-velocity carve-out for cases where outputs occur faster than per-output human review, such as algorithmic trading or automated cyber response. In those cases, the checkpoint moves upstream: the named human signs the policy that authorizes

the class of outputs, and the four channels attach to that signature. Speed moves the checkpoint but does not remove it.

The Constitutional Wall. The wall is the point where AI assistance becomes unauthorized delegation. The Named-Human Test is the Constitutional Wall stated as a question. Any reduction in substantive human engagement at a checkpoint converts AI Governance back to Responsible AI, regardless of whether a human is physically present. A checkpoint is where a named human exercises binding authority that cannot be delegated to the AI being governed, and the standard is that binding authority, not mere presence.

Individual accountability survives audit. The test rests on a real legal pattern. In regulated domains, liability and audit attach to named individuals when their contribution is substantive. PCAOB Rule 3502, amended in 2024, holds a named associated person liable when they negligently, directly, and substantially contribute to a firm violation. Sarbanes-Oxley Sections 302 and 906 require named officers to personally certify. FDA 21 CFR 11.100 requires electronic signatures unique to one identity-verified individual, non-reusable and non-reassignable. The UK ICO requires that a human reviewer have the authority and competence to change the decision, and the EDPB-endorsed WP29 guidelines require meaningful human intervention by someone able to change the outcome. These precedents distinguish binding authority from nominal presence in practice and answer the rubber-stamping objection: a name on a form without authority, competence, evidence access, and a demonstrable record of override does not pass.

Enacted-law validation. The Named-Human Test is no longer only a synthesis of analogous regimes. It now substantially tracks the same operational standard in new AI-specific law. Colorado SB 26-189, signed May 14, 2026, defines meaningful human review in statute as review by an individual designated by the deployer who has authority to approve, modify, or override a consequential decision. Practitioner analysis of the enrolled text describes additional behavioral requirements: that the reviewer is trained, considers primary evidence, does not default to the system output, and has access to the information needed to understand it. That analysis states the anti-rubber-stamping reading directly: a reviewer who ratifies a system output without genuine deliberation fails the standard. The UK Data (Use and Access) Act 2025 reaches the same place from the opposite political direction, making meaningful human involvement, a person with the authority and competence to change the decision, the hinge of its automated-decision provisions. This is among the strongest external validations because it is AI-specific, enacted in 2026, and reproduces the binding-authority and anti-rubber-stamping elements together.

Dissent preserved. The objection that naming a human can create a scapegoat without real power, the moral crumple zone, is real. Some regimes attach liability to the organization or the operator instead of a named individual, and defenders of those models point to the EU civil-liability proposal and Federal Reserve model-risk guidance as alternatives.

Practitioners who implement the Named-Human Test do not deny collective accountability. They add the named-individual obligation on top, as SOX does, and the test fails wherever the named human lacks the actual institutional power to halt the system.

The counter-risk to collective accountability is now documented at scale. The MIT Delphi study identifies what it calls an accountability sink: "responsibility shared across many actors becomes responsibility held by none." When 272 experts independently identify that distributed responsibility without named-individual binding produces a governance vacuum, they are identifying the same failure pattern the Named-Human Test is designed to prevent. The test exists because organizational accountability, without a named person who answers through at least one of the four channels, diffuses into a structure where no one is accountable when the system produces harm.

The central equivalence

Different fields describe the same line, who answers for the output, and that line is Axis B: the named human. The equivalence is a routing rule, and it must be stated as a conditional, never as an identity.

The same agentic capability routes to one side or the other depending on a single variable. Without a named human bound to the output, the system routes left: Artificial Intelligence in the Tale of Two AIs, Responsible AI (automation without named-human authority) in the three-outlook taxonomy, factory mode in the plain metaphor. With a named human bound to the output, the same system routes right: Augmented Intelligence, AI Governance, handmade mode.

This is a routing rule, and the conditional framing is what keeps the framework from collapsing the two axes. Stating the equivalence as an identity (saying that agentic AI "is" Responsible AI) would collapse the two axes the framework exists to separate, because agentic capability is a property on Axis A while Responsible AI is a category on Axis B. They co-occur on the left only when no named human is bound, and the moment an organization binds a named human to the outputs, the same capability moves right without any change to the underlying technology.

The framework unifies prior work because each existing framework resolves to the same question. The three AI outlooks (Ethical AI, Responsible AI, AI Governance) resolve to whether a named human holds binding authority. The Tale of Two AIs (Artificial versus Augmented) resolves to the same variable. Factory versus Handmade, Checkpoint-Based Governance, the Constitutional Wall, and the HAIA-RECCLIN method can all be routed through one operational test: is a named human bound?

Augmented Intelligence and AI Governance are not synonyms. They co-occur on the right side, but they answer different questions. Augmented Intelligence is the collaboration mode, what the human and AI do together. AI Governance is the authority layer, who holds binding power and answers for the result. The relationship is dependency: augmentation is only held on the right by governance. A system can attempt augmentation with no named human, and the moment it does, it collapses back to the left and becomes Responsible AI no matter how collaborative it looks. Without binding authority, collaborative tools revert to unsupervised automation.

Finding 1: The record

Several distinct signals confirm that the replacement agent is losing acceptance, and each must be kept separate by scope so none is overstated.

Among agentic AI specifically, Gartner predicted in June 2025 that over 40 percent of agentic AI projects will be canceled by end of 2027, citing cost and unclear value alongside inadequate risk controls. The same release forecasts 15 percent of day-to-day work decisions made autonomously and 33 percent of enterprise software with agentic AI by 2028. This is perfect dissent inside one source: Gartner supports both the pressure and the opportunity. The cancellation signal is a forecast, not capability death.

Across broader AI initiatives (not agents exclusively), S&P Global Market Intelligence reports that 42 percent of companies abandoned most of their AI programs in 2025, up from 17 percent in 2024. This is realized behavior, not a forecast, and it is the stronger acceptance-failure evidence among the market signals. RAND cites estimates that more than 80 percent of AI projects fail and documents recurring root causes through interviews. These figures cover enterprise AI broadly, not agentic deployments in isolation.

On agent-specific benchmarks, Carnegie Mellon's TheAgentCompany study shows the best agents completing 24 percent of common office tasks autonomously, with partial-credit scoring reaching roughly 30 percent. Salesforce's CRMArena-Pro shows about 58 percent accuracy on simple single-step tasks falling to about 35 percent on multi-step operations, corroborating the CMU result. These are production-maturity signals that confirm a gap between capability demonstration and reliable deployment, and they do not prove the capability is dead.

The deployment-governance gap is the sharpest signal. EY reports that 76 percent of surveyed companies are using or planning to use agentic AI within the next year, while only about a third have proper protocols across EY's Responsible AI framework, and 56 percent are familiar with the associated risks. An EY follow-up survey reported widespread AI-related financial losses, with the vast majority of respondents experiencing losses and few

identifying the correct controls. Deployment intent is outrunning control and accountability, and that gap is a governance failure, not a capability judgment.

Scoped to generative AI broadly (not agents specifically), MIT NANDA's *The GenAI Divide* (2025) reports that 95 percent of enterprise generative AI pilots delivered no measurable financial return.

The deployment scale is visible at the infrastructure level. Cloudflare reports that automated traffic, driven increasingly by agentic AI, surpassed human HTTP traffic in 2026, with the split reaching 57.5 percent bot to 42.5 percent human. The total includes traditional crawlers alongside agentic agents, but Cloudflare's CEO attributed the crossover to agentic traffic growing faster than forecast, arriving over a year ahead of his 2027 prediction. The agents within that traffic are doing bounded tasks at massive volume: checking prices, comparing flights, scraping content. Cloudflare measures HTTP requests, not engagement or consequential decisions, but the scale confirms that factory-mode deployment is already the dominant pattern on the open internet.

Experts have now quantified the risk. In a June 2026 MIT Delphi study, 272 international AI specialists judged 18 of 24 AI risk domains as having more than 10 percent probability of catastrophic outcomes, defined as more than one million deaths or more than $100 billion in financial loss, within the next five years under business-as-usual conditions. Even under a scenario where pragmatic mitigations are implemented, those experts still judged all 24 risks as retaining more than 5 percent catastrophic probability, with five remaining above 10 percent. Their assessment, built through three iterative rounds, confirms that the deployment-governance gap the market signals identify carries catastrophic tail risk that current governance structures are not designed to absorb.

The scope condition applies here, and it runs in both directions. The technology has not delivered reliable human replacement: agents that complete a quarter of tasks and collapse on multi-step operations are useful tools under supervision, not the autonomous replacements the era promised. At the same time, markets are rejecting the replacement premise through realized cancellation and abandonment, not waiting for the technology to mature. The failure is both functional and social, and the two reinforce each other. Agentic capability under human authority survives, but the project of replacing humans with unchecked autonomous agents does not.

Finding 2: The social movement

The political backlash against AI concentrates on replacement, and the replacement frame drags everything else with it.

More than 1,000 AI-related bills were introduced in 2025, mostly at the US state level. That figure measures legislative attention, not public opinion directly, and the US legislative picture is now bifurcating (see Finding 3 below). Yet the direction of the attention is consistent: disclosure and transparency requirements recur across the bills, with human oversight as the recurring structural demand. The backlash is bending corporate deployment toward accountability, partly ahead of regulation.

Investigative reporting supplies the documented grievances the backlash runs on. Karen Hao's *Empire of AI* documents data-annotation labor conditions, data-center power and water use, and research-ecosystem capture. Five of her nine major claims held under independent testing, corroborated by Rest of World, Time, New York Magazine, Reuters, and Bloomberg, and the one figure she corrected is excluded here. Her colonial-empire framing is her thesis, attributed as interpretation. Tristan Harris reaches a mass audience with the structural diagnosis that the AI race repeats the attention economy's profit-over-safety failure. His argument is labeled diagnosis of the dynamic, not evidence for a factual claim.

Differential polling shows that public rejection concentrates on unaccountable consequential decisions, not automation as such. Pew finds 71 percent oppose AI making final hiring decisions (2023) while views of AI reviewing applications are more mixed. Grant Thornton finds only 5 percent of organizations allow agents to execute high-stakes decisions without human review.

The counterevidence is stated openly: the public does not reject all automation. Harvard Business School research (Friis and Riley, 2025) finds people supported automating roughly 30 percent of occupations at current capability, rising to 58 percent under more advanced AI, with about 12 percent drawing moral resistance regardless of capability. Acceptance coexists with rejection, divided by the consequential-decision line.

That baseline captures 2025, but by early 2026, the accountability line had hardened considerably. An ITIF/Morning Consult poll finds 79 percent insisting a human must make the final decision before any use of lethal force, with 75 percent saying AI is not yet reliable enough for life-or-death military decisions without human oversight. In a Future of Life Institute survey of likely voters, 80 percent support human oversight with clear limits and corporate accountability, and 77 percent insist AI remain under human control with authority to stop systems when needed, while only 10 to 11 percent favor lightly regulated development. An Economist/YouGov poll from May 2026 records 71 percent of Americans stating AI development is moving too fast, with pessimists outnumbering optimists roughly

two to one. Respondents across these surveys concentrate their rejection on the same variable: consequential decisions where no named human answers for the result. In 2025, people showed support rising with capability. By 2026, that support stops at the checkpoint where accountability would otherwise disappear.

The public draws the accountability line consistently across surveys. Whether this reflects a stable preference or a reaction to the industry's failure to distinguish replacement from augmentation is not yet directly tested. The two-axis separation offers the vocabulary regardless of which mechanism proves correct.

Finding 3: The regulation (as of May 2026)

The human-review principle persists across jurisdictions and is in places hardening into enacted definition. The direction, however, is contested and bifurcating across at least three poles, and regulatory convergence is not the claim.

The EU: the principle holds while deadlines shift. EU AI Act Article 14 requires effective human oversight by natural persons for high-risk systems. The capacities it lists are operational: understand the system and monitor it, remain aware of automation bias, correctly interpret output, decide not to use or to disregard, override, or reverse the output, and intervene or stop the system. Article 14 does not enumerate four oversight models (human in command, in the loop, on the loop, over the loop); those come from earlier EU ethics guidance, not from the Article 14 text. This is override authority held by a person, and it validates the Named-Human Test. GDPR Article 22(3) supplies breadth beyond the high-risk taxonomy by granting a right to obtain human intervention in automated decisions with legal or similarly significant effects, applied extraterritorially, answering the objection that Article 14 is relevant only to high-risk systems.

The EU AI Act's prohibited-practice penalties (up to EUR 35 million or 7 percent of global turnover) have applied since February 2, 2025. The high-risk obligations were originally set for August 2, 2026. European legislators then moved to defer those deadlines: the Digital Omnibus on AI, provisionally agreed May 7, 2026, pushes them to December 2, 2027 for stand-alone Annex III systems and to August 2, 2028 for AI embedded in regulated Annex I products. Regulators have not yet formally adopted the Omnibus; adoption is expected before August 2, 2026, and until publication in the Official Journal the original Act remains the baseline and August 2, 2026 stays a live date. Legislators deferred dates, but they did not remove Article 14.

The US: deregulatory at the federal level, still legislating at the state level. The United States is moving away from the EU high-risk model, but the narrower human-review principle survives. Colorado passed the first comprehensive state AI law (SB 24-205) in 2024, then repealed and replaced it with SB 26-189, signed May 14, 2026, while SB 24-205's

enforcement was stayed. The new law drops risk-management programs and impact assessments, along with the duty of care against algorithmic discrimination. It replaces them with a narrower automated-decision-technology regime that still requires notice, adverse-outcome disclosure, data correction, and meaningful human review and reconsideration. In parallel, xAI sued to enjoin the original law, the US Department of Justice intervened in support (the first federal move against a state AI law), and enforcement was stayed. Executive Order 14365 sets a national policy against a 50-state patchwork, and Texas scaled its TRAIGA proposal back from an EU-modeled bill toward a narrower harmful-use prohibition.

The meaningful human review that Colorado retained is defined in the enrolled statute as review by an individual designated by the deployer who has authority to approve, modify, or override a consequential decision. Practitioner analysis of the enrolled text describes the reviewer as trained, considering primary evidence, and not defaulting to the system output. That substantially matches the Named-Human Test, including the anti-rubber-stamping wall, written into US law.

The US picture is not uniform deregulation. While the federal government pushes anti-patchwork preemption, states keep enacting AI laws. Connecticut's SB 5, signed into law by Governor Lamont on May 27, 2026, creates requirements for frontier AI developers, AI companions, automated employment decision technology, and content provenance, with effective dates staggering from October 2026. California finalized automated-decision-technology rules in 2025, and both states proceeded despite the federal posture, which means the bifurcation runs inside the US, not only across the Atlantic.

The UK: a third pole. The UK Data (Use and Access) Act 2025 (Section 80, new Article 22A UK GDPR) relaxes the EU-style prohibition on solely-automated decisions for non-sensitive data, moving the UK away from the EU model. That is a genuine liberalization, not merely a persistence of prior rules. Yet the Act turns its automated-decision provisions on whether there is meaningful human involvement: active involvement by a person with the authority and competence to change the decision. The UK both liberalizes the prohibition and centers the exact concept the Named-Human Test rests on.

The international baseline. The Council of Europe Framework Convention on Artificial Intelligence, opened for signature September 5, 2024, is the first legally binding international AI treaty. It was signed by the EU, the US, the UK, Israel, and others. Its principles cover transparency and accountability, with oversight provisions and remedies for affected persons and a required independent oversight mechanism per party. Two caveats are mandatory: it is technology-neutral and human-rights-framed, not a named-individual-per-decision mandate, and the US signature is signed but not ratified, made under the prior administration, and now in visible tension with the current federal posture.

Even where regulation retreats from prescriptive high-risk regimes, the narrower principle of meaningful human review for consequential decisions persists on both sides of the Atlantic and is in places hardening into enacted definition. This persistence confirms the Named-Human Test is tracking what the regulatory environment considers non-negotiable, but the direction is contested and bifurcating, and regulatory convergence is not the claim.

This section carries a 30-to-60-day expiry. Items to re-verify before publication: the Omnibus formal adoption date, the xAI v. Weiser litigation posture under the new Colorado law, and whether the Council of Europe treaty has moved toward US ratification. All dates in this section are current as of late May 2026.

Finding 4: The definitional collapse

The industry defines "agentic" by minimal supervision, placing the standard agent in Responsible AI (automation without a named human bound to the output) by construction. Vendors and standards bodies helped normalize the collapse, and it is the mechanism through which the market sells an autonomous factory as if it were handmade.

A nine-platform adversarial review plus independent verification against the source texts confirms that NIST AI RMF, ISO/IEC 42001, the OECD AI Principles, the Microsoft Responsible AI Standard, and Google's AI principles all express human oversight as role-based or organizational accountability. On their face, these frameworks generally permit role-based or organizational accountability and do not clearly require a named natural person with binding authority over a specific consequential output. ISO/IEC 42001 is role-based at the standard level (Clause 5.3, Annex A.3); the named-individual reading appears only in audit-practice guidance layered on top of the standard. Industry practice is beginning to narrow this gap, with leading implementations moving toward a single accountable owner model, and the stipulative definition used here describes the current dominant mode rather than a permanent feature.

OpenClaw makes the collapse concrete. A local agent framework that connects models, skills, files, and online services answers the process-control question on Axis A without answering the accountability question on Axis B. OpenClaw can read messages, send emails, access calendars, and act across everyday digital services, and the word "local" can sound like control. But local execution answers where the process runs, and the Named-Human Test asks who holds binding authority over what the process does. Security research has already moved toward checkpoints, interruption mechanisms, and human confirmation for agent runtimes, precisely because broad permissions without governance create system-level exposure that the agent's capability alone cannot address.

The competing research around OpenClaw's social network ecosystem illustrates the attribution problem the definitional collapse enables. Some researchers treat agent-only

activity as evidence of autonomous social behavior, while others find that the most viral claims of emergence were overwhelmingly human-influenced. The governance lesson from that dispute is the same distinction the two axes draw: the question is who authorized it and who answers for it, and if no one can answer, the system is Responsible AI regardless of how autonomous it appears.

The collapse is connected to a broader pattern. When the marketing of AI systems strips the words that distinguish replacement from augmentation, the public draws the accountability line by rejecting systems where no one answers for the output. The two-axis separation is the vocabulary that the definitional collapse has withheld.

Finding 5: The mechanism

Why did deployment outrun governance? The question matters because the answer determines whether the gap is a temporary learning curve or a structural feature.

The Economic Override Pattern, first described in *Governing AI: When Capability Exceeds Control* (Puglisi, 2025, Chapter 2), proposes the structural answer: deployment incentives override governance regardless of institutional intent, because speed and cost efficiency outweigh safety and accountability when the cost of ungoverned deployment is not priced in. Karen Hao's reporting that capability selection at the major labs is revenue-driven provides investigative corroboration drawn from internal documents and roughly 300 interviews, confined here to the five claims that held under independent testing. Tristan Harris's widely-heard argument that the AI race structurally overrides safety for capability is the same pattern stated as diagnosis. Both are cited as support for the published Economic Override Pattern, not as its origin and not as independent proof.

Verified incentive-misalignment evidence separates the pattern from ordinary adoption lag. Grant Thornton's 2026 AI Impact Survey finds 46 percent of leaders say AI underperforms because controls and compliance are not working, while only 11 percent identify governance as the function most needing focus, and 78 percent lack confidence they could pass an independent AI governance audit within 90 days. Organizations know the control gap and under-invest in it, which is incentive misalignment, not a neutral learning curve. Databricks states the gap directly: when organizational incentives reward shipping models fast, teams treat governance as a blocker. The speed of deployment has outrun even the infrastructure layer's predictions: Cloudflare's CEO expected the human-to-bot traffic crossover in 2027, but automated traffic driven by agentic growth surpassed human HTTP requests over a year ahead of that forecast.

The MIT Delphi study provides the broadest independent corroboration to date. Across 272 experts, competitive dynamics ranked among the top five severity risks, and their language reproduced the pattern without citing it: "any individual developer that slows down to

invest in safety bears a direct competitive cost while the safety benefits accrue to society at large," and "the actors judged most responsible have structural reasons not to act on that responsibility." The study also identified that the two actors experts named as most responsible, general-purpose AI developers and governance actors, "are therefore at risk of weakening together rather than compensating for one another" under competitive pressure. That is consistent with the Economic Override Pattern stated by an independent academic panel, not a single author, and it separates the pattern from adoption lag the same way the testable prediction does: by showing that the gap is driven by incentive structure, not by insufficient time.

The pattern generates a testable prediction that distinguishes it from ordinary lag. If economic override drives the gap, high-competition sectors should show measurably wider deployment-governance gaps and higher post-deployment cancellation than low-competition sectors in the same window. Ordinary adoption lag predicts uniform delays governed by internal capacity, not by competitive pressure. The prediction is stated as a hypothesis for future testing.

The Economic Override Pattern is classified as a Tier 2 working concept: supported by observable evidence and a testable prediction, not independently validated as formal theory. The Tier 2 label is a disclosed limitation. Hao's colonial-empire framing is her thesis, not proof, and Harris is diagnosis, not data, and the causal gap between the pattern and ordinary adoption lag is not yet closed.

Finding 6: The resolution

The separation of the two axes, and the Named-Human Test that operationalizes Axis B, does three jobs at once.

For regulators, it supplies the operational line that survives even where prescriptive high-risk regimes are rolled back. The meaningful-human-review principle persists in Colorado, the UK, and the EU because it tracks the axis that regulatory systems independently concluded is non-negotiable: not how autonomous the system is, but whether a named human answers for the output.

For enterprises, it supplies a deployable test that is also a compliance test. Factory or handmade is a procurement question that any officer can answer, and the same test produces a regulatory-readiness answer. It preserves factory mode as a legitimate choice where stakes allow, because the framework is classification, not prohibition. The high-velocity carve-out shows what governed agentic deployment looks like in practice: an algorithmic trading firm whose named compliance officer signs the policy authorizing a class of automated trades, with the four channels of accountability attaching to that signature, is AI Governance even though no human reviews each trade. Speed moved the

checkpoint upstream, but a named human is still bound. That is the difference between an agent running free and an agent running governed. The same logic applies to consumer-facing agents: OpenClaw governed by restricted permissions, logged actions, and a named human with approval authority over consequential outputs is AI Governance in this framework. OpenClaw operating across email, calendar, files, and external services without meaningful checkpoint authority remains Responsible AI in factory mode, regardless of how personal or local it feels.

For the public and the political sphere, it supplies the distinction between AI that replaces and AI that answers to a human. Agentic AI was always going to fail as a replacement project because the technology could not reliably replace the humans it promised to displace, and the public rejected the premise before the technology could catch up. Both failures point to the same resolution. The distinction between agents under human authority and agents running free, once named, is durable because it tracks what the evidence and the polling both confirm: accountability for the result is what people and institutions care about, and capability alone does not satisfy it.

The replacement paradigm is not the only position the evidence weakens. The prohibition position, Stop AI, fails for the same reason major prohibitions of persistent capability have failed throughout history: capability that exists does not become uninvented because a governing body forbids it. Alcohol prohibition drove production underground and produced a 400 percent increase in methanol poisoning mortality because clandestine operations drop safety protocols to remain invisible. The 1990s encryption export controls pushed cryptographic development offshore without improving security. Drug prohibition repeated the same dynamic across decades and continents. In each case, the technology persisted, the safety infrastructure disappeared, and the enforcement institutions themselves became targets for capture. Banning AI development removes governance from the systems that continue to operate, which is the opposite of the outcome the prohibition is designed to achieve.

The regulation-only position also fails when it stops at principles without binding named-human authority. Ethics boards that meet quarterly, safety teams that lack authority to block deployments, and published principles without implementation protocols that enable verification are governance theater, and the Economic Override Pattern predicts exactly why. When competitive pressure makes safety costly and enforcement mechanisms lack teeth, organizations optimize for the constraints they face, not the ones they have promised to observe. Centralized control without individual accountability creates capture points: three ratings agencies with concentrated authority over structured finance amplified systemic risk through correlated failures in 2008, and any global AI authority built on the same architecture would face the same vulnerability. Responsible AI (automation without a named human bound to the output) has a legitimate place for bounded, reversible

applications, but it cannot substitute for governance, and selling it as governance is the mislabeling documented across the six findings above.

The pattern is visible even at the frontier labs with the strongest published safety commitments. An earlier analysis documented that Anthropic's 80-page Constitution for Claude comprehensively addresses character formation and internal safeguards while leaving external governance architecturally undefined: no binding checkpoint authority, no named human accountable through four channels, no mechanism that survives the removal of voluntary cooperation (Puglisi, "The Missing Governor," January 2026). Six months later, Anthropic published internal data showing that over 80 percent of its production code is now authored by Claude, with Claude reviewing Claude's code and the company describing the human role as "narrowing at each step." That is machine checks machine at production scale, which is Responsible AI by the three-tier definition, regardless of the sophistication of the constitution that governs the machine's character.

That leaves the resolution the evidence supports: Augmented Intelligence under AI Governance, meaning structured human-AI collaboration where a named human governs the method, holds binding authority at the checkpoint, and answers for the result through the four accountability channels. The collaboration arc that spans human cognitive history, from spoken language through writing through printing through the internet, has required the same response at every transition: more structured, more accountable, more governed forms of collaboration. History is supposed to teach us what works, and every medium that introduced new categories of error eventually demanded that response. The governance architecture followed because the alternative was compounding failure until an institutional crisis forced the correction. The question the evidence leaves is why, knowing that, we are treating AI as the exception instead of building upon what we have already learned.

Method

The thesis originated with the author, built from prior published work on checkpoint-based governance, the three-tier AI framework, and the Named-Human Test. A formal mapping documented the hypothesis, scope conditions, evidence requirements, and constraints before any prose was written. The mapping was the skeleton; the paper was written from it, not the reverse.

Source research began with parallel dispatch across multiple independent AI platforms, each tasked with locating, verifying, and stress-testing the evidence base. Where platforms returned conflicting results, the conflict was investigated against primary sources through live-web verification. Two citations identified as fabrications during this process were quarantined and never entered the paper.

The first complete draft was submitted for independent prepublication review across eleven AI platforms simultaneously, each receiving the same prompt and the manuscript with no guidance on what to find. The reviews were synthesized by a twelfth platform serving as independent auditor. Convergent findings drove revisions; divergent findings were investigated and resolved through primary-source checking.

The author then introduced evidence and sources that no AI platform had surfaced independently: the Cloudflare traffic crossover data, the Anthropic recursive self-improvement disclosure tied to a January 2026 governance-gap analysis the author had published, the MIT Delphi study of 272 experts, the 2026 public backlash polling, and two previously published articles supplying the prohibition and governance-theater arguments the subtitle promises. One additional source, the OpenClaw agent framework, was introduced through a platform review and developed by the author into a case study. Each addition was tested through a subsequent review round before entering the paper.

The paper then went through multiple rounds of author-directed challenges: an editorial precision pass, two independent fact-verification rounds, and a full source-validation audit across three platforms. The four accountability channels were revised at the canon level during this process when the author identified that the original term did not accurately describe the mechanism it was meant to capture. Every substantive decision throughout this process was made by the author. The AI platforms executed research, drafted prose, and tested claims. The author governed what the paper says.

References

European Commission, Council, Parliament. (2026). *Digital Omnibus on AI*, provisional agreement, 7 May 2026: Annex III deferred to 2 December 2027, Annex I to 2 August 2028, pending formal adoption.

European Union. (2024). *Regulation (EU) 2024/1689 (AI Act), Article 14: Human Oversight.* Official Journal.

European Union. (2016). *Regulation (EU) 2016/679 (GDPR), Article 22: Automated individual decision-making*.

Colorado General Assembly. (2024). *SB 24-205, Colorado AI Act*. And (2026) *SB 26-189*, signed 14 May 2026, repealing and replacing SB 24-205; effective 1 January 2027; retains meaningful human review.

Connecticut General Assembly. (2026). *SB 5, Connecticut Artificial Intelligence Responsibility and Transparency Act*. Signed 27 May 2026; effective dates stagger from 1 October 2026.

Council of Europe. (2024). *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law* (CETS 225, opened for signature 5 September 2024).

United Kingdom. (2025). *Data (Use and Access) Act 2025*, Section 80 (new Article 22A UK GDPR).

Puglisi, B. C. (2025). *Governing AI: When Capability Exceeds Control*, Chapter 2 (Corporate Incentives and Economics). ISBN 9798349677687.

Gartner. (25 June 2025). *Over 40% of Agentic AI Projects Will Be Canceled by End of 2027.*

S&P Global Market Intelligence. (2025). 42% of companies abandoned most AI initiatives, up from 17% in 2024.

EY. (June 2025). *Responsible AI Pulse* (76% using or planning agentic AI; about one-third with proper controls; 56% familiar with risks; 975 leaders, 21 countries).

MIT NANDA. (2025). *The GenAI Divide: State of AI in Business 2025* (95% of generative AI pilots without measurable ROI).

Carnegie Mellon University. (2025). *TheAgentCompany* (arXiv 2412.14161; best agents 24 to 30% task completion).

Salesforce. (2025). *CRMArena-Pro* (58% simple, 35% multi-step).

RAND. (2024). *The Root Causes of Failure for AI Projects* (RR-A2680-1; cites estimates over 80% fail).

Harvard Business School. (2025). Friis and Riley, *Performance or Principle* (WP 26-017; 30% of occupations supported for automation at current capability, 58% under advanced AI, 12% moral resistance).

Pew Research Center. (2023). AI-in-hiring polling (71% oppose AI making final hiring decisions).

ITIF/Morning Consult. (February 2026). Survey of 1,976 U.S. adults on AI in military operations. 79% insist human must make final decision before any use of lethal force; 75% say AI not reliable enough for life-or-death military decisions. https://itif.org/publications/2026/02/26/survey-most-americans-say-tech-companies-should-allowed-set-ai-limits/

Human Statement / Future of Life Institute. (March 2026). Survey of 1,004 likely U.S. voters (Feb 19-20, 2026). 80% support human oversight; 77% insist AI remain under human control. Advocacy-commissioned polling. https://humanstatement.org/poll-americans-support-pro-human-principles/

Economist/YouGov. (May 2026). 71% of Americans say AI development is moving too fast.

Grant Thornton. (2026). *AI Impact Survey* (46% blame governance; 11% prioritize it; 78% cannot pass audit; only 5% allow fully autonomous high-stakes without review).

Cloudflare / Prince, M. (June 2026). Automated bot traffic surpassed human HTTP traffic: 57.5% bot vs. 42.5% human. Total includes traditional crawlers and agentic agents; CEO attributed crossover to agentic growth arriving ahead of 2027 forecast.

Hao, K. (2025). *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. Five of nine major claims held under independent testing (Puglisi, *Empire of Evidence*, 2026).

Harris, T. (November 2025). Center for Humane Technology. Structural diagnosis of AI race dynamics on The Diary of a CEO.

PCAOB. (2024). Rule 3502, amended 12 June 2024. Named associated person liability for negligent, direct, substantial contribution to firm violations. https://pcaobus.org/news-events/news-releases/news-release-detail/pcaob-modernizes-its-rules-by-strengthening-accountability-for-contributing-to-firm-violations

Sarbanes-Oxley Act, Sections 302 and 906. Named officer personal certification of financial statements.

FDA. 21 CFR 11.100. Electronic signatures unique to one identity-verified individual, non-reusable and non-reassignable.

UK ICO. Meaningful-human-intervention guidance for automated decision-making.

EDPB/WP29. (2018). WP251, Guidelines on Automated Individual Decision-Making and Profiling. Meaningful human intervention by someone able to change the outcome.

14 CFR 91.3. Pilot in command authority and responsibility in aviation.

NIST. (2023). *AI Risk Management Framework 1.0* (AI RMF). https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

ISO/IEC. (2023). *42001:2023, Artificial Intelligence Management System*. Role-based accountability at standard level; named-individual reading in audit-practice guidance.

OECD. (2019, updated 2024). *OECD AI Principles*. https://oecd.ai/en/principles

Microsoft. *Responsible AI Standard, v2*. Organizational accountability framework.

Google. *AI Principles* (revised 2025). Organizational accountability framework.

IBM. (2026). *What is Agentic AI?* and *What Are AI Agents?*

xAI v. Weiser (D. Colo., filed 9 April 2026); US DOJ intervention (24 April 2026); enforcement stayed (27 April 2026). Executive Order 14365.

Blum, D. (2010). The Poisoner's Handbook: Murder and the Birth of Forensic Medicine in Jazz Age New York. Penguin. See also: Methanol poisoning during Prohibition, documented in PMC/NIH archives. https://pmc.ncbi.nlm.nih.gov/articles/PMC2972336/

Electronic Privacy Information Center. (1999). *Cryptography and Liberty: An International Survey of Encryption Policy*. (1990s encryption export controls drove development offshore without improving security.)

Financial Crisis Inquiry Commission. (2011). *The Financial Crisis Inquiry Report*. (Three ratings agencies with concentrated authority over structured finance amplified systemic risk through correlated failures.)

Puglisi, B. C. (January 2026). The Missing Governor: Anthropic's Constitution and Essay Acknowledge What They Cannot Provide. basilpuglisi.com. https://basilpuglisi.com/the-missing-governor-anthropics-constitution-and-essay-acknowledge-what-they-cannot-provide/

Anthropic Institute. (June 2026). When AI builds itself: Our progress toward recursive self-improvement, and its implications. https://www.anthropic.com/institute/recursive-self-improvement

Saeri, A. K., Graham, J., Noetel, M., Slattery, P., Thompson, N., et al. (June 2026). Prioritization of Risks from Artificial Intelligence: A Delphi Study of 272 International Experts. MIT AI Risk Initiative and MIT FutureTech. Data: https://osf.io/pj2qr

IEEE Spectrum. (2026). Moltbook, the AI Agent Network, Heralds a Messy Future. https://spectrum.ieee.org/moltbook-agentic-ai-agents-openclaw

OpenClaw. (2026). Personal AI assistant framework. https://openclaw.ai/ and https://github.com/openclaw/openclaw

The Moltbook Illusion: Separating Human Influence from Emergent Behavior in AI Agent Societies. (2026). arXiv:2602.07432 (15.3% autonomous, 54.8% human-influenced; no viral phenomenon originated from a clearly autonomous agent).

Podcast: "The Other AI: Audio Briefings on Augmented Intelligence and AI Governance"

Available on Spotify, Apple Podcasts, Amazon Music, and YouTube