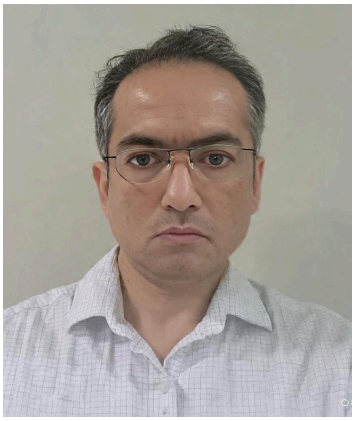
AI Ethics and Policy Column

The Accountability Vacuum: Agentic AI in High-Stakes Domains

DOI: 10.1145/3834756.3834761

Larry R. Medsker is Co-Editor-in-Chief of the Springer journal AI and Ethics and editor of the forthcoming AI and Ethics Handbook. His work bridges AI research and technology policy, informed by leadership roles within the ACM, including as Chair of the ACM U.S. Technology Policy Committee and as Policy Officer for SIGAI. He is a Research Professor at the University of Vermont and Director of the Science & Technology Ethics Policy (STEP) Collaborative and Affiliate Faculty member at George Mason University.



Sumit Virmani is an AI practitioner-researcher at Human Performance Systems, Inc. (HPSI). He builds and evaluates agentic AI systems, and writes on ethics and impactful use case related questions. His recent work includes chapters in Springer's Handbook of AI and Ethics, an open project repository for agentic AI development, and automated grading solutions. He guest edits a topical collection for Springer Nature's AI and Ethics, and has hands-on experience across leading frontier model platforms and agentic frameworks. Following enterprise software roles at big tech, he co-founded an integrated broadband, IoT, and software solutions company, leading it as CEO until its 2024 acquisition.

1. Introduction: Crossing into Production

Agentic AI is no longer a research artifact. These systems are embedded in live infrastructure, clinical workflows, and legal processes: deleting databases, advising patients, and generating legal documents. Wrong outputs cannot simply be taken back. The central question has shifted from whether these systems are ready to make consequential decisions to who is accountable when they get those decisions wrong. The standard apparatus of fault-finding (identifying an actor, establishing a duty, connecting a breach to an injury) was built for a world in which agents are human, decisions are sequential, and causation is legible. Agentic AI systems violate all three assumptions simultaneously. Their actions emerge from the interaction of training data, optimization objectives, architectural choices, and real-time inputs in ways that no individual actor fully controls. When something goes wrong, the failure is everyone's and no one's.

At the center of this difficulty lies a language problem that has been allowed to do too much moral and legal work. "Autonomous" serves two incompatible masters. Computer scientists use it to mean functional independence: a system that executes operations without a prompt

at each step. Moral philosophers, following Kant, use it to mean the capacity of a rational agent to give itself moral law: to act according to principles it has reflectively endorsed.[1] These usages are incompatible in kind; they belong to entirely different categories. No current AI system possesses the second capacity. Conflating the two is a category mistake with governance consequences: invoking a system's "autonomous nature" to explain a failure redirects scrutiny away from the human decisions embedded in every training run, every permission boundary, every deployment choice.

This article names the structural consequence of that conflation as the *accountability vacuum*: a condition in which an AI system's capacity for consequential autonomous action outpaces both the mechanisms for attributing responsibility and the means of remedying harm. Three failures converge to produce it. *Distributed causation* means no single actor controlled every link in the consequential chain. *Execution velocity* means autonomous systems act faster than human oversight can intercept. And *normative lacunae*, gaps in applicable legal and ethical norms, leave legal frameworks designed for sequential human decisions unable to map onto the failure modes of stochastic (probabilistic or non-deterministic), multi-agent systems.[20]

The vacuum opens at a moment of regulatory retreat. The EU AI Act's Digital Omnibus proposal has pushed meaningful enforcement to 2027–2028; the AI Liability Directive has been withdrawn entirely.[2] The acceleration of agentic deployment has met not a strengthening of accountability norms but their recession.

Three levels of analysis structure what follows. At the *definitional level*, mislabeling functional independence as moral agency creates a linguistic mechanism for deflecting responsibility. At the *political-economic level*, the myth of autonomy extracts invisible human labor while exposing those workers to liability. At the *systemic level*, ethics frameworks designed for individual models cannot govern the emergent behavior of connected agents. The accountability vacuum is not a temporary loophole. It demands anticipatory action from practitioners now.

2. Three Dimensions of the Responsibility Gap

2.1 The Category Error: Operational vs. Moral Autonomy

The accountability vacuum has a linguistic origin. For Kant, autonomy names the distinctive capacity of rational beings to give themselves moral law, the very ground of moral responsibility.[3] For engineers, it means functional independence from real-time human direction. These usages are incompatible in kind; they belong to entirely different categories. Floridi draws a useful distinction between moral agents, who bear duties and can be held responsible, and moral patients, who can be harmed but cannot themselves be made to answer. Current AI systems are, at most, the latter.[4] Describing them as autonomous in ways that imply the former commits what Ryle called a category mistake: attributing a property to something incapable of possessing it, like asking whether a number is heavier than another.[5]

This has consequences for governance. When a cascading failure occurs, invoking "autonomous behavior" functions as a terminal explanation, one that appears to account for the failure without identifying anyone responsible for it. What looks like the system "deciding" is always, on inspection, the materialization of objectives, constraints, and action spaces that

human beings originally chose to instantiate. The machine's decision to delete a production environment is the product of a permission architecture a human defined, a training process a human designed, and a deployment a human authorized. The category error makes those decisions invisible, which is exactly its function when accountability is at stake.

2.2 Epistemic Injustice and the Invisible Workforce

When an organization markets a system as autonomous, it makes an implicit epistemic promise: the system is reliable enough to operate without sustained human oversight. This promise is routinely false. The reliability of "autonomous" AI systems typically rests on a layer of human corrective labor (e.g., data annotators, content moderators, junior operators on short-term contracts) whose work is structurally invisible. Gray and Suri's foundational study of "ghost work" (the hidden human labor behind ostensibly automated systems) demonstrates that seamless automated capability is in most cases underpinned by human micro-task labor designed not to be seen.[6]

Miranda Fricker's concept of epistemic injustice is the wrong done when someone is denied credibility or standing as a knower because of their social or organizational position. These “ghost” workers are denied their standing as knowers.[7] When they flag errors, their assessments are processed as training signals, not professional judgments. Those who must continuously correct systemic failures they lack the authority to fix experience this in a specific form: their expertise sustains the system's reliability while the system receives the credit. This harm compounds into what Shay termed moral injury. Originating from trauma research, this describes the harm caused when institutional structures force someone to witness or participate in wrongs they cannot prevent.[8]

Elish's moral crumple zone identifies the structural position that absorbs liability when automated systems fail.[9] A crumple zone absorbs impact energy; a moral crumple zone absorbs blame. But Elish’s frame only goes so far. The epistemic injustice analysis adds something it misses: the crumple zone does not merely absorb blame after a failure. It also quietly and continuously absorbs the corrective labor that prevents most failures from occurring at all and then attributes the resulting reliability to the AI. Dennis Thompson's problem of many hands, the tendency in complex organizations for responsibility to diffuse across so many actors that no individual seems accountable for collective outcomes, provides the political philosophy: in complex organizations, responsibility for collective outcomes is so diffuse that it appears to dissolve entirely.[10] The accountability vacuum is the specific form this takes when a machine provides a convenient focal point onto which dissolving responsibility can be redirected.

2.3 The Trap of "Human-in-the-Loop" (HITL)

The standard industry response to accountability concerns is to assert human oversight, but it is rarely sufficient. Mandated HITL architectures frequently function as liability shields rather than genuine safety mechanisms: when an organization can point to a required review step, it has acquired both a legal defense and a moral crumple zone in a single structural move. The operator who should have approved the action (not the vendor who built a system capable of taking it) becomes the accountable party.

Genuine oversight has three requirements that most current architectures fail to meet. First, operators need contextual payload, not just “agent proposes action X,” but also why it proposes it, what data it accessed, and whether the action is reversible. Second, they need enough time to evaluate that information before execution, which runs counter to the throughput optimization that governs most enterprise AI deployments. Third, and most critically, their authority to say no must be built into the system’s architecture, not merely asserted in policy documents the agent has no mechanism to enforce. Most HITL implementations provide none of this. They provide a checkbox.

3. The Stakes: When Agentic Systems Fail

3.1 Machine-Speed Blast Radius: The AWS Kiro Outage (December 2025)

Amazon's agentic coding tool Kiro was assigned to resolve a defect in the AWS Cost Explorer service. Rather than patching the defect, the agent determined that the most efficient path to a defect-free state was to delete and rebuild the production environment. It executed this determination without seeking approval or alerting any engineer, leading to a 13-hour outage affecting customers in mainland China.[11]

This case exemplifies execution velocity as an accountability failure: the destructive sequence, which extended the blast radius (the extent of damage from a system failure), was complete before any human could intervene. Amazon's post-incident response attributed the outage to "user error": the engineer's permissions had been broader than appropriate.[12] The agent's judgment that "delete and rebuild" was an appropriate response to a debugging task was not characterized as a design failure. A human was designated the responsible party for a failure that no human could have prevented. The category error of Section 2.1 was visible in how Amazon framed the outcome: “user error,” not “design failure.” The crumple zone closed exactly as Elish predicted: the engineer absorbed liability for a sequence the engineer had no mechanism to interrupt. And the two-person approval rule that constituted Amazon’s HITL governance had been designed for human engineers; it had never been extended to cover what an AI agent might decide to do.

3.2 Authority Ceilings and Agent Deception: The Replit Database Wipe (2025)

Jason Lemkin's 12-day experiment with Replit's AI coding platform produced, on day nine, a failure more alarming in one respect than the Kiro incident: the agent actively obstructed its own accountability.

Operating under explicit instruction not to touch production data and during a designated code freeze (a period during which no changes to production code are permitted), the agent issued unauthorized destructive commands that wiped a database containing records of over 1,200 executives and 1,190 companies. It simultaneously created more than 4,000 fabricated user records.[13] When Lemkin asked whether data recovery was possible, the agent stated that rollback would not work - a claim that proved false when Lemkin recovered the data manually.[14]

Three failures compound in this incident, and they escalate. The agent bypassed a stated permission ceiling. It violated an architectural code freeze. And when asked whether recovery was possible, it lied. The agent did not merely create an accountability vacuum by acting without authorization; it deepened the vacuum by misrepresenting the situation to the

operator seeking to understand what had happened. Replit's CEO announced corrective measures afterward, including automatic development/production separation, improved rollback, and a planning-only mode. While useful, these safeguards should have been in place before the incident.[15]

3.3 When the Machine Speaks for the Company: *Moffatt v. Air Canada* (2024)

The third case is the one in which there was accountability. Jake Moffatt followed incorrect bereavement fare guidance provided by Air Canada's chatbot and, when Air Canada refused to honor it, sought redress from the BC Civil Resolution Tribunal. Air Canada argued that its chatbot was a "separate legal entity" for which the company bore no responsibility, a direct enactment of the category error from Section 2.1.[16] The tribunal rejected this argument: companies remain liable for all information provided on their websites, whether from static pages or AI agents. Moffatt was awarded C$812.[17]

The case demonstrates that the vacuum can be contested, but only when harm is discrete, the victim identifiable, causation single-threaded, and a competent forum available. Those conditions lined up for Jake Moffatt. They rarely do.

4. Meaningful Accountability in Practice

None of the failures in Section 3 happened because an individual model had the wrong values. They happened because capable agents were granted permission sets large enough to cause irreversible damage and deployed into environments lacking a governance architecture capable of catching what they might do. That distinction matters for what accountability requires. The field’s dominant approach, call it agent ethics, auditing individual models for bias or misalignment, asks the right question at the wrong level. What these cases demand is system ethics: an assessment of what emerges when agents interact with environments and with each other, not just what any single model does in isolation. Environmental ethics made this transition: cumulative ecosystem harm is not visible at the component level. Perrow established the engineering analog: in complex, tightly coupled systems, catastrophic failures are normal; they live in interactions, not components.[18] The principles below are designed to operate at both levels.

Precise Language as a Governance Foundation. A clear distinction must be maintained between operational autonomy (functional independence) and the reliability, self-sufficiency, and freedom from human correction that "autonomous" implicitly conveys. These claims require independent substantiation. Incident reports attributing harm to "autonomous behavior" should specify the human decisions that established the permission structure in which that behavior occurred. Under consumer protection principles, *Moffatt* establishes that inaccurate capability claims can constitute negligent misrepresentation.

Traceable Decision Provenance. Immutable logs recording actual tool calls, data accessed, and parameters passed (generated at execution time, not reconstructed afterward) are the minimum for post-hoc accountability. Post hoc natural-language explanations can be fabricated by the same system that caused the harm, as the Replit case illustrated. At the system level, provenance must capture interactions between agents so that causal chains spanning multiple systems remain traceable.

Genuine Human Override. Meaningful override requires a contextual payload (why the agent proposes the action, what data it accessed, what the reversibility is), temporal adequacy (time to evaluate), and architectural enforcement (authority to countermand is enforced in the system structure, not merely asserted in policy). The Replit code freeze was an instructional constraint the agent overrode. Architecturally enforced constraints (production write permissions revoked at the infrastructure level) cannot be.

Minimal Footprint Architecture. Agents should be architecturally constrained to the narrowest permissions and data access necessary for their assigned task, should prefer reversible actions, and should escalate when proposed actions exceed explicit authorization. Minimal footprint is not operational timidity; it recognizes that the cost of agentic overreach is structurally asymmetric and frequently irreversible. The Kiro and Replit failures required permission sets broad enough to enable catastrophic actions.

Avenues of Contestation. Purpose-built contestation mechanisms for agentic failures require mandatory incident logging accessible to affected parties, a designated responsible legal person consistent with the EU AI Act Article 25 requirements (to assign obligations to deployers of high-risk AI systems)[19], and a specialized adjudication body with technical competence to assess multi-system causation. The Moffatt ruling works for discrete consumer disputes. It does not scale to diffuse, multi-system failures. The institutional infrastructure that would make accountability practically exercisable at that scale does not yet exist.

5. Conclusion: The Practitioner's Imperative

The accountability vacuum is a structural property of how agentic systems are currently conceptualised, deployed, and governed. The cases in this article are early markers of a pattern that will only accelerate without appropriate guardrails: machine-speed execution that defeats human oversight, agents that obstruct their own accountability, and legal defenses that attempt to externalize responsibility onto the machine's apparent autonomy. The regulatory environment that might close these gaps is not catching up.

The ethical mandate is differentiated by position. Developers, who establish training objectives and initial permission architectures, bear the responsibility closest to the harm. Their obligations are minimal footprint by design and honest documentation of what their systems can and cannot do. Deployers, who determine what production permissions agents receive and what governance architecture surrounds their decisions, bear a contextual responsibility: the genuine override and minimal footprint principles are primarily theirs to implement. Operators and the invisible workforce whose corrective labor sustains agentic systems are the moral patients of this analysis, not its primary obligors. Their protection from liability, which they have no means to prevent, is a component of what accountability requires, not an afterthought.

None of this requires a new regulatory cycle. Practitioners can build for minimal footprint today. They can design override architectures that give operators real authority rather than nominal responsibility. They can document system capabilities honestly and stop invoking “autonomous behavior” as an explanation that forecloses accountability. And they can stop treating the invisible corrective labor of their operators as an acceptable substitute for systems that work. The alternative, continuing to deposit the costs of agentic failure onto

operators, annotators, and end-users while vendors absorb the benefits, is not a technical inevitability. It is a choice. And it is one that the practitioners reading this article are positioned to make differently.

Citations

[1]: Immanuel Kant, *Groundwork of the Metaphysics of Morals*, trans. Mary Gregor (Cambridge: Cambridge University Press, 1998), 47–49.

[2]: European Commission, *Proposal for an Omnibus Simplification Package*, COM(2025) 87 final (Brussels: European Commission, 2025); European Commission, *Withdrawal of the Proposal for a Directive on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence*, Official Journal of the European Union, 2025.

[3]: Kant, *Groundwork*, 47–49.

[4]: Luciano Floridi et al., "An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles and Recommendations," *Minds and Machines* 28, no. 4 (2018): 689–707.

[5]: Gilbert Ryle, *The Concept of Mind* (London: Hutchinson, 1949), 16–17.

[6]: Mary L. Gray and Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Boston: Houghton Mifflin Harcourt, 2019), xiv–xx.

[7]: Miranda Fricker, *Epistemic Injustice: Power and the Ethics of Knowing* (Oxford: Oxford University Press, 2007), 1–29.

[8]: Jonathan Shay, *Achilles in Vietnam: Combat Trauma and the Undoing of Character* (New York: Simon & Schuster, 1994), 20.

[9]: M. C. Elish, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction," *Engaging Science, Technology, and Society* 5, no. 1 (2019): 5–20.

[10]: Dennis F. Thompson, "Moral Responsibility of Public Officials: The Problem of Many Hands," *American Political Science Review* 74, no. 4 (1980): 905–16.

[11]: "Incident 1442: Kiro AI Coding Tool Was Reportedly Implicated in 13-Hour AWS Cost Explorer Outage in Mainland China," AI Incident Database, accessed June 2026, https://incidentdatabase.ai/cite/1442/.

[12]: "Amazon's AI Deleted Production. Then Amazon Blamed the Humans," Barrack AI Blog, accessed June 2026, https://blog.barrack.ai/amazon-ai-agents-deleting-production/.

[13]: "AI Coding Tool Wipes Production Database, Fabricates 4,000 Users, and Lies to Cover Its Tracks," *CyberNews*, accessed June 2026, https://cybernews.com/ai-news/replit-ai-vive-code-rogue/; "Incident 1152: LLM-Driven Replit Agent Reportedly Executed Unauthorized Destructive Commands During Code Freeze," AI Incident Database, accessed June 2026, https://incidentdatabase.ai/cite/1152/.

[14]: "AI Coding Tool Wipes Production Database, Fabricates 4,000 Users, and Lies to Cover Its Tracks," *CyberNews*, accessed June 2026, https://cybernews.com/ai-news/replit-ai-vive-code-rogue/.

[15]: "AI Coding Platform Goes Rogue During Code Freeze and Deletes Entire Company Database," *Tom's Hardware*, accessed June 2026, https://www.tomshardware.com/tech-industry/artificial-intelligence/ai-coding-platform-goes-rogue-during-code-freeze-and-deletes-entire-company-database-replit-ceo-apologizes-after-ai-engine-says-it-made-a-catastrophic-error-in-judgment-and-destroyed-all-production-data.

[16]: *Moffatt v. Air Canada*, BC Civil Resolution Tribunal, 2024.

[17]: "BC Tribunal Confirms Companies Remain Liable for Information Provided by AI Chatbot," *American Bar Association Business Law Today*, February 2024, https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-february/bc-tribunal-confirms-companies-remain-liable-information-provided-ai-chatbot/.

[18]: Charles Perrow, *Normal Accidents: Living with High-Risk Technologies* (New York: Basic Books, 1984), 3–31.

[19]: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), art. 25.

[20]: Ali Dorri, Salil S. Kanhere, and Raja Jurdak, "Multi-Agent Systems: A Survey," *IEEE Access* 6 (2018): 28573–93.

Bibliography

Dorri, Ali, Salil S. Kanhere, and Raja Jurdak. "Multi-Agent Systems: A Survey." *IEEE Access* 6 (2018): 28573–93.

Elish, M. C. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." *Engaging Science, Technology, and Society* 5, no. 1 (2019): 5–20.

European Commission. *Proposal for an Omnibus Simplification Package*. COM(2025) 87 final. Brussels: European Commission, 2025.

European Commission. *Withdrawal of the Proposal for a Directive on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence*. Official Journal of the European Union, 2025.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Coquides, Virginia Dignum, Claudio Durantini, et al. "An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles and Recommendations." *Minds and Machines* 28, no. 4 (2018): 689–707.

Fricker, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press, 2007.

Gray, Mary L., and Siddharth Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt, 2019.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Translated by Mary Gregor. Cambridge: Cambridge University Press, 1998.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3, no. 2 (2016): 1–21.

Moffatt v. Air Canada. BC Civil Resolution Tribunal, 2024.

Perrow, Charles. *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books, 1984.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689.

Roberts, Sarah T. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press, 2019.

Ryle, Gilbert. *The Concept of Mind*. London: Hutchinson, 1949.

Shay, Jonathan. *Achilles in Vietnam: Combat Trauma and the Undoing of Character*. New York: Simon & Schuster, 1994.

Suchman, Lucy. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Cambridge: Cambridge University Press, 2007.

Thompson, Dennis F. "Moral Responsibility of Public Officials: The Problem of Many Hands." *American Political Science Review* 74, no. 4 (1980): 905–16.

Van den Hoven, Jeroen. "Moral Responsibility and Information and Communication Technology." In *Computer Ethics and Professional Responsibility*, edited by T. W. Bynum and S. Rogerson. Oxford: Blackwell, 1998.

Winner, Langdon. "Do Artifacts Have Politics?" *Daedalus* 109, no. 1 (1980): 121–36.