Learn from the blame game when AI causes harm

Jay Killoran & Andrew Park

2026

Peter B. Gustavson School of Business

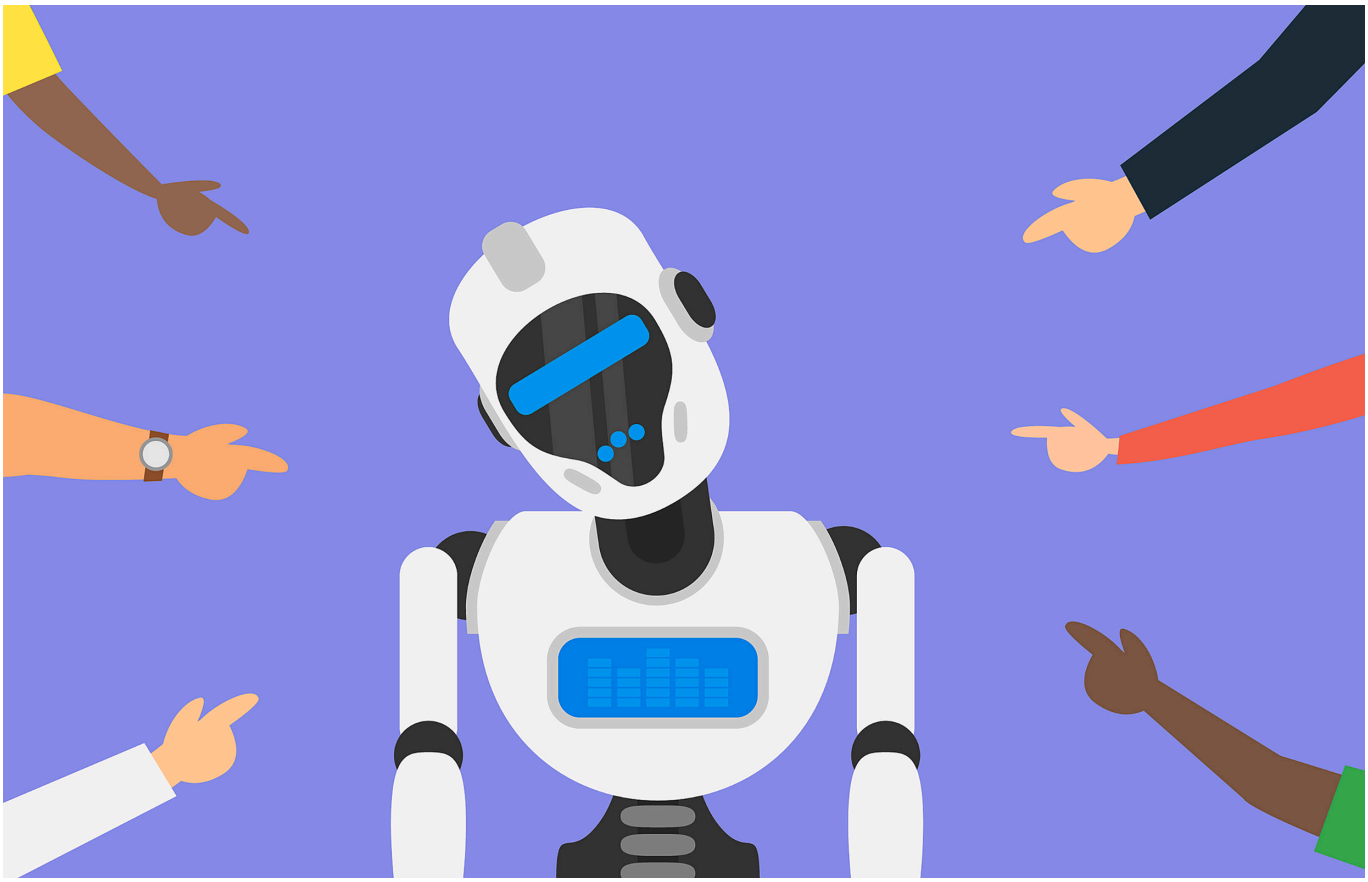
Faculty Publications

Original citation:
Killoran, J., & Park, A. (2026). Learn from the blame game when AI causes harm. *Proceedings of the National Academy of Sciences - PNAS*, *123*(17), Article e2528408123. https://doi.org/10.1073/pnas.2528408123

PNAS OPINION

Learn from the blame game when AI causes harm

Jay Killoran[a,1] and Andrew Park[a]



When AI systems cause harm, people instinctively blame the technology–a reaction often dismissed as scapegoating. While AI systems cannot be morally responsible, blaming AI can be informative if treated as a diagnostic signal. Image credit: Shutterstock/Tarikdiz

In 2023, families of deceased Medicare Advantage patients filed a class action lawsuit against UnitedHealth Group, alleging that an AI algorithm, nH Predict, systematically denied medically necessary post-acute care (1). Physicians had approved the care, but nH Predict denied it. The lawsuit alleged that UnitedHealth knew the tool had a 90% error rate, yet continued to deploy it, counting on fewer than 1% of patients appealing. For its part, UnitedHealth has denied wrongdoing, maintaining that nH Predict was never used to make coverage determinations but, rather, served only as a clinical guide and that all coverage decisions were made by human reviewers (2). The case, which is ongoing, ignited a wider debate: Who was to blame? The engineers? The company? The case managers? Or the AI system?

This case illustrates a recurring dilemma in AI governance. As large language models (LLMs) and autonomous decision-making systems (together here called AI systems) become embedded in high-stakes domains such as healthcare, criminal justice, and financial services, they have been implicated in harms. These include wrongful arrests (3), contributions to teenage suicide (4), and denial of medically necessary care. While this case involves a predictive algorithm, similar dynamics of public blame, diffuse responsibility, and institutional evasion are increasingly visible in harms involving agentic AI—systems to which humans delegate decision authority across consequential tasks, from LLM-generated medical misinformation (5) to AI coding agents causing infrastructure outages (6).

When these systems produce harm, the public seeks someone or something to blame. And blame is often treated as a problem to be managed rather than a diagnostic signal we can learn from. This is a mistake. Here, we break down this argument into two points. First, we reconceptualize public blame toward AI systems as diagnostic

Author affiliations: [a]Gustavson School of Business, University of Victoria, Victoria, BC V8N 4V3, Canada

Author contributions: J.K. and A.P. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: jkilloran@uvic.ca.

Published April 22, 2026.

data about socio-technical failure. Second, we offer a three-part taxonomy—Ignorance, Diminished Control, and Condemnation—as an analytical instrument for scholars, practitioners, and policymakers when they respond to AI-related harms.

Blame Game

Social science scholarship has historically interpreted blame toward AI systems as scapegoating: AI becomes a target onto which society projects anxieties, deflecting responsibility from humans and institutions (7–9). This framework emerged in the era of deterministic, rule-based IT systems, when investigators could identify bugs and trace causal chains. Social scientists agreed that attributing fault to such systems, rather than designers, in this way was a flawed and unethical deflection.

Modern AI systems are different. They are complex architectures, often with opaque mechanisms that make it difficult to trace outcomes to specific design decisions (10). This opacity need not render responsibility ambiguous. In fields such as aviation and pharmaceuticals, full traceability is rare, yet responsibility can be clearly attached to decisions made by people during design, deployment, and governance. The same applies to AI.

Still, opacity does make blame more socially fraught and more likely to land on a visible target such as the AI system rather than underlying institutions. This tendency is reinforced by research showing that people perceive modern AI systems to have human-like agency (11), making them compelling targets for blame.

Continuing to view blame toward AI systems as unhelpful scapegoating is now insufficient. Scapegoating theory—which holds that blame is sometimes displaced onto less culpable or less powerful targets in order to protect more powerful actors from scrutiny—rightly warns against this kind of deflection. In the AI context, this means resisting the tendency to treat the algorithm as the responsible party, thereby shielding the engineers, executives, and institutions whose decisions actually shaped the system's behavior. However, this theory dismisses the useful possibility that public blame can help in understanding what went wrong. We argue that while AI systems cannot be held responsible, blaming them can be productive if reframed as informative telemetry. Blame can then be interpreted as a diagnostic signal: evidence of where socio-technical systems have failed and a guide toward the network of users, designers, and institutions that produced the harm. This can then shape accountability, remediation, and governance.

Scapegoating occurs when blame is displaced onto less powerful targets, preserving the status of more powerful actors (7–9). With AI, scholars argue that blaming these systems obscures human culpability across design, deployment, and regulation. Treating AI as autonomous allows humans to absolve themselves while maintaining the myth of technological objectivity.

This dynamic is dangerous, as AI systems scale across high-stakes domains. When an AI-assisted clinical tool generates harmful recommendations, institutions may deflect by suggesting "the algorithm made the decision," obscuring human judgment embedded in data, design, and deployment (12). When predictive policing targets marginalized communities, responsibility is offloaded onto technology, insulating structural inequities (13). And when insurers deploy AI to override physician determinations, as alleged against UnitedHealth, responsibility diffuses across developers, administrators, and automated processes, complicating accountability. In each case, no actor is clearly responsible, and nothing changes.

The Dangers of Scapegoating

Scapegoating theory has three limitations when it comes to AI. The first is overcorrecting against blame. Scapegoating implies that blaming AI is irrational. Yet, blame reflects real perceptions of agency, opacity, and unpredictability that modern AI systems produce (11, 14). Dismissing it overlooks its communicative function.

The second is neglecting the symbolic dimension. Blame signals violated expectations, breached trust, and harm requiring recognition. Treating it as deflection ignores its value for learning and investigation.

The third is that scapegoating fails to provide alternatives. The theory offers little guidance on how institutions should interpret or respond to public reactions to AI harm. Holding AI responsible for mistakes or harm is misguided. Responsibility demands moral agency: the capacity for intention, moral understanding, and repair (15, 16). AI systems lack these capacities. Whether predictive models or LLMs, they optimize mathematical functions and probabilities, not moral values (17). The appearance of agency in modern systems makes this easy to overlook, but appearance is not agency. Confusing the two permits institutional evasion. Treating AI as responsible absolves designers and institutions. And "autonomous decision-making" becomes a defense, creating an accountability vacuum where harms persist.

AI can mimic intentional action through reasoning and planning (18), but this is not equivalent to human agency. Treating it as such conflates function with moral capacity. And framing AI as responsible blurs the line between tool and agent, distorting expectations and weakening governance discourse (19).

If AI systems cannot be responsible, what should we make of the human tendency to blame them anyway? Blame has always carried informational value. It signals violated expectations and accountability gaps. This is especially important in AI, where opacity, distributed responsibility, and rapid deployment outpace governance. Treating blame as telemetry—data about system failure—provides a key interpretive move, and one missing in current academic and policy debates.

Taxonomy of Blame

To demonstrate this, we draw on prior empirical work (20–22) to propose a taxonomy of blame narratives: Ignorance, Diminished Control, and Condemnation. Each signals a different failure, implicates different actors, and suggests different responses.

Blame narratives rooted in ignorance arise when people attribute harmful outcomes to AI systems because they do not understand how those systems function, make decisions, or generate outputs. The system is treated as a black box, and blame emerges from opacity rather than demonstrable malfunction. Ignorance shows itself in expressions of confusion ("I don't know how it works"), uncertainty about causal mechanisms ("the system just did it"), or appeals to hidden flaws ("something must be wrong inside the algorithm"). When

ignorance narratives dominate public reactions to an AI-related harm, the implication is that there is a transparency or explainability failure in the system's design or deployment, and the corrective response should target those gaps directly.

Blame framed as diminished control arises when AI systems constrain human agency, positioning the technology as usurping an individual's authority or decision-making role. Signs include concerns about being overruled ("I was ignored"), claims of inability to intervene ("I couldn't change what it decided"), or complaints of forced compliance ("the system wouldn't let me do otherwise"). This narrative type has become particularly salient in the era of agentic AI systems that execute multistep tasks autonomously, such as AI-assisted clinical decision support (5). With diminished control narratives, the implication is that the system's design has inappropriately shifted decision-making authority away from human actors and that governance or design interventions are needed to restore meaningful human oversight.

Condemnation arises when AI systems are blamed for outcomes that violate ethical or safety expectations. Unlike ignorance and diminished control, condemnation reflects a judgment that a normative boundary has been breached. Reactions include perceived violations of moral standards ("the AI is unethical"), direct attributions of harm ("the system put lives at risk"), or appeals to justice ("it should never allow that to happen"). Condemnation narratives are particularly informative because they reveal not just that something went wrong, but that something went wrong in a way that people regard as fundamentally unacceptable. The implication is that the system's design, training, or deployment has enabled an outcome that violates shared social norms; the corrective response must engage not just technical fixes, but the normative frameworks that govern the system's deployment and use.

Taking Responsibility

Together, these narratives form a diagnostic repertoire. Ignorance signals knowledge gaps; diminished control signals disempowerment; condemnation signals normative breach. They also map onto actors: designers, deployers, institutions, and regulators. This mapping transforms blame into a governance resource. By asking what blame reveals about system failure and distributed responsibility, institutions can investigate root causes, rather than symptoms. Courts do not ask whether AI is responsible; they ask which human actors made which decisions under what obligations. This is precisely the question that our diagnostic taxonomy is designed to answer.

Liability frameworks already distribute responsibility across these actors. Designers and developers fall under product liability (23). Deploying institutions fall under negligence and duty of care (24). Users fall under contributory negligence (25). Our blame narratives map onto these categories. Ignorance signals failures in transparency and safeguards. Diminished control signals institutional overreliance on automation. Condemnation signals broader normative failures engaging multiple forms of liability.

Recent cases reflect this shift. In the litigation against UnitedHealth, plaintiffs allege deployment of a model with a known error rate to deny care (1), engaging multiple liability regimes. Regulatory investigations increasingly focus on design, deployment, and governance decisions, rather than system behavior alone (24).

For scholars, this alignment suggests an emerging research agenda to link blame narratives to legal outcomes and accountability. For practitioners, it reframes public blame as a governance signal, rather than a communications problem. For policymakers, recurring blame patterns indicate systemic regulatory gaps.

AI governance will not be achieved by resolving whether AI is responsible. It will be achieved by building systems that convert human reactions—including blame—into structured accountability. The reframing we offer here complements existing legal frameworks by ensuring that public reactions are analyzed and acted upon, rather than dismissed. And as AI systems continue to integrate into our society, better analysis of blame can help work out why things go wrong and guide us toward the right lessons to learn.

1. E. Napolitano, UnitedHealth uses faulty AI to deny elderly patients medically necessary coverage, lawsuit claims. *CBS News* (2023). https://www.cbsnews.com/news/unitedhealth-lawsuit-ai-deny-claims-medicare-advantage-health-insurance-denials/. Accessed 2 March 2026.
2. C. Ross, B. Herman, UnitedHealth faces class action lawsuit over algorithmic care denials in Medicare Advantage plans. *Stat* (2023). https://www.statnews.com/2023/11/14/unitedhealth-class-action-lawsuit-algorithm-medicare-advantage/. Accessed 24 March 2026.
3. K. Hill, Wrongfully accused by an algorithm. *NY Times*, 24 June 2020. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html. Accessed 7 October 2025.
4. K. Roose, Can A.I. be blamed for a teen's suicide? *NY Times*, 23 October 2024. https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html. Accessed 8 October 2025.
5. D. Wu *et al.*, First, do NOHARM: Towards clinically safe large language models. arXiv [Preprint] (2025). https://arxiv.org/abs/2512.01241 (Accessed 2 March 2026).
6. J. Morales, AWS outages caused by AI coding bot blunder, report claims. *Tom's Hardware*, 20 February 2026. https://www.tomshardware.com/tech-industry/artificial-intelligence/multiple-aws-outages-caused-by-ai-coding-bot-blunder-report-claims-amazon-says-both-incidents-were-user-error. Accessed 1 March 2026.
7. H. Nissenbaum, Computing and accountability. *Ethics Inform. Tech.* **1**, 351–358 (1994).
8. M. Bauer, J. Cahlíková, J. Chytilová, G. Roland, T. Želinský, Shifting punishment onto minorities: Experimental evidence of scapegoating. *Econ. J.* **133**, 1626–1640 (2023).
9. S. C. Florman, *Blaming Technology: The Irrational Search for Scapegoats* (St. Martin's, New York, 1981).
10. A. Asatiani *et al.*, Sociotechnical envelopment of artificial intelligence: An approach to organizational deployment of inscrutable artificial intelligence systems. *J. Assoc. Inf. Syst.* **22**, 325–352 (2021).
11. B. S. Vanneste, P. Puranam, Artificial intelligence, trust, and perceptions of agency. *Acad. Manage. Rev.* **50**, 726–744 (2026).
12. J. M. Logg, J. A. Minson, D. A. Moore, Algorithm appreciation: People prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* **151**, 90–103 (2019).
13. K. Alikhademi *et al.*, A review of predictive policing from the perspective of fairness. *Artif. Intell. Law* **30**, 1–17 (2022).
14. H. Altehenger, L. Menges, The point of blaming AI systems. *J. Ethics Soc. Philos.* **27**, 287–314 (2024).
15. K. Gray, L. Young, A. Waytz, Mind perception is the essence of morality. *Psychol. Inq.* **23**, 101–124 (2012).
16. B. F. Malle, S. T. Magar, M. Scheutz, "AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma" in *Robotics and Well-Being*, M. I. Aldinhas Ferreira *et al.*, Eds. (Springer Nature Switzerland, 2019), pp. 111–133.
17. P. B. Rainey, M. E. Hochberg, Could humans and AI become a new evolutionary individual? *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2509122122 (2025).
18. S. Lebovitz, H. Lifshitz-Assaf, N. Levina, To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organ. Sci.* **33**, 126–148 (2022).
19. T. Birkstedt, M. Minkkinen, A. Tandon, M. Mäntymäki, AI governance: Themes, knowledge gaps and future agendas. *Internet Res.* **33**, 133–167 (2023).
20. J.-F. Bonnefon, I. Rahwan, A. Shariff, The moral psychology of artificial intelligence. *Annu. Rev. Psychol.* **75**, 653–675 (2024).
21. J. A. Killoran, "Blame ascriptions toward autonomous agents," Doctoral dissertation, Queen's University, Kingston, Canada, QSpace and Library and Archives Canada (2025).
22. R. M. McManus, C. C. Mesick, A. M. Rutchick, Distributing blame among multiple entities when autonomous technologies cause harm. *Pers. Soc. Psychol. Bull.* **51**, 2068–2084 (2025).
23. M. C. Buiten, Product liability for defective AI. *Eur. J. Law Econ.* **57**, 239–273 (2024).
24. P. Panarese, M. M. Grasso, C. Solinas, Algorithmic bias, fairness, and inclusivity: A multilevel framework for justice-oriented AI. *AI & Soc.*, 10.1007/s00146-025-02451-2 (2025).
25. A. D. Selbst, Negligence and AI's human users. *Boston Univ. Law Rev.* **100**, 1354–1360 (2020).