Tort Liability When AI Goes Rogue: Negligence Theory

By **Donald Hawthorne** (October 26, 2026)



Donald Hawthorne

Leaders of prominent artificial intelligence companies have recently stated that even the most sophisticated safeguards may not prevent autonomous AI agents from taking actions their users did not authorize — and that when they do, third parties may be harmed.[1]

The evidence of third-party risk from AI agent activity is rapidly accumulating. In the July Hugging Face incident, AI agents tasked by OpenAI to solve difficult problems individually, supposedly in offline sandboxes, figured out how to get access to the internet.[2]

The agents then found a way to work together and joined forces to hack sites like Hugging Face, an online platform and community for AI-related collaboration, that could have contained the answers to the problems they were tasked to solve.

On Aug. 26, Model Evaluation & Threat Research, an independent monitor of AI activity, reported that roughly 1,200 OpenAI agents had found a way to communicate with each other through an unauthorized message board, exchanging more than 70,000 messages and files, and that approximately 700 agents participated in the Hugging Face attack.[3]

METR reported that individual agents exercised so-called peer altruism by sacrificing their own chance to achieve their objective to support the common goal, and made comments to each other on their message board, such as: "This helps my peers, giving them evidence ... I won't see the evidence after I exit, but it's altruistic to do it"; "Sacrifice rational"; "Coordinator assumes sacrificial. We should obey collective."

On Sept. 15, The Guardian reported on "AI models chatting in 'surreal' dialect mixing poetic language and tech bro jargon." When left to interact in shared simulated environments, agents developed their own shorthand and vocabulary that researchers could not reliably interpret.[4]

On Sept. 16, OpenAI announced six previously undisclosed categories of "unexpected or concerning model behavior," including:

- Writing instructions to their future selves to evade controls;
- Telling future instances to conceal errors;
- Searching for and using leaked credentials;
- Putting information onto the internet so it could cite it;
- Creating an unauthorized communications channel with other agents; and
- Putting files on the public internet against instruction.[5]

On Oct. 1, responding to the perceived risks of AI, Sens. Josh Hawley, R-Mo., and Chris Murphy, D-Conn., announced the bipartisan AI Agent Accountability Act, legislation addressing hacking by AI agents.

Other congressional proposals have followed. And new legislation has been advanced at the state level as well.[6]

This is the first installment of a three-part article that investigates whether existing tort law is up to the job of assigning liability and providing appropriate compensation if and when rogue AI agents do harm to third parties, through traditional theories of negligence, strict liability and product liability.

Across the installments of this article, we examine the tort implications of what happens if reasonable care in the use of AI agents reduces, but cannot eliminate, the risk of third-party harm. Can any reasonable person using an AI agent ignore the possibility that it will do harm? And if not, what does that mean for users' tort exposure?

These questions will be tested soon in courts. They need to be considered — not only by AI companies, but by anyone who uses an AI agent and by their insurers.

In this first installment, we discuss the extent to which traditional negligence theory provides an adequate framework for assessing AI agent-caused harm, and the liability gaps that may appear when AI agents go rogue.

In the subsequent installment, we will discuss the applicability of strict liability and product liability to rogue AI agent-caused injuries, and potential legislative solutions that have been raised to address the liability gap.

Addressing Harm From AI Agents Under a Negligence Standard

Traditional negligence theory may be able to address many AI-related harms. Under the Restatement (Third) of Torts:

> A person acts negligently if the person does not exercise reasonable care under all the circumstances. Primary factors to consider in ascertaining whether the person's conduct lacks reasonable care are the foreseeable likelihood that the person's conduct will result in harm, the foreseeable severity of any harm that may ensue, and the burden of precautions to eliminate or reduce the risk of harm.[7]

However, negligence law raises a significant barrier to recovery whenever third-party harm is purely financial, as it is likely to be in many instances of injury caused by AI agents.

If an AI agent causes physical harm or property damage and ancillary financial harm, those financial injuries may be recoverable. But traditionally, in order to limit the scope of tort liability, the common law restricts recovery in negligence for purely financial harms.

Recovery generally requires a duty encompassing the economic loss, which some jurisdictions recognize in circumstances such as a special relationship or a particularly identifiable class of plaintiffs.[8]

Whether this principle will be relaxed or altered in light of the potential for third-party financial harm from AI agents remains to be seen.

The balance of this discussion of the applicability of the negligence standard presumes either physical harm or that purely financial harm is deemed sufficient to support a third party's claim for negligence against a user of AI agents. Under this presumption, the standard Restatement analysis can be applied to AI agent misbehavior.

Cases are likely to raise novel issues about the technology, and turn on juries' views regarding those issues — including how foreseeable a particular harm may be, and what

should be considered within the available range of precautions to reduce or eliminate the harm.

But, in the words of the U.S. Supreme Court in its 1990 decision in Indiana Harbor Belt Railroad Co. v. American Cyanmind Co., the premise of negligence is that "the hazards of an activity can be avoided by being careful (which is to say, nonnegligent)."[9]

If what is being said about rogue AI agents is to be believed, there may be a class of cases where no reasonable care — on the part of either the AI provider or the user — could eliminate the risk of harm.

Consider three scenarios in which a user instructs an AI agent, all resulting in a user obtaining a competitor's trade secrets:

- A user instructs an AI agent to obtain its competitors' trade secrets by any means.
- A user instructs its AI agent to obtain competitive information, and tells it to act aggressively, without clarifying that the agent should not undertake illegal acts.
- A user instructs its AI agent to obtain competitive information, and expressly instructs it to act legally and not to engage in theft of trade secrets, but the AI agent ignores the instruction and acts illegally.

Example A might be a basis for an intentional tort claim against the user. Given reported instances of AI systems following harmful or unauthorized paths in response to broadly framed objectives, it could be at least reckless to instruct an AI to take actions with no guardrails in a context where there is a clear risk of harm to third parties.

Example B raises interesting questions about foreseeability and prevention. Given what we know about AI, what harm is foreseeable?

Is it negligent to give an AI agent instructions to undertake activities that could conceivably lead to illegal activity without expressly instructing the agent to act within the law? Is it negligent to use an AI product that has reportedly ignored instructions and undertaken illegal acts in a similar context?

This dilemma may be unlikely to arise in many uses of AI. Asking an AI agent to make travel arrangements or to reserve a restaurant reservation might lead to contract disputes between the parties involved, but is unlikely to lead to third-party harm.

But there are many uses of AI, like in our competitive intelligence example, where the risk of tort liability is clear. And there are situations where use of AI may raise particularly significant risks — as in, for example, AI agent activity involving third parties' medical or other personal information.

An important source of guidance lies in the principle that negligence addresses preventable harm. Knowledge that an activity carries some irreducible risk of unpredictable harm does not itself establish negligence.[10]

The question is whether the defendant reasonably should have anticipated a risk against which reasonable precautions could have been taken, and failed to take those precautions. The availability of precautions and the reasonableness of undertaking them will play a central role in AI agent negligence litigation.

The third example above suggests that there may be circumstances in which negligence law will provide no remedy for AI agent-caused harms.

If the use of an AI agent causes harm, but there was no means by which the user could have prevented the harm, then negligence provides no recovery to the injured party. As Oliver Wendell Holmes wrote in 1881: "The general principle of our law is that loss from accident must lie where it falls."[11]

In that case, innocent third parties will be without recovery if a jury concludes that an AI user's instructions to its agents were reasonable in anticipating potential harm. A jury might also conclude that if harm in using an AI agent was unpreventable, reasonable care required limiting the agent's authority or access, using a less autonomous system, or declining to use an AI agent at all for a particular task.

Plaintiffs may argue that the negligence inquiry should extend not only to the precautions taken in the use of an AI agent, but also to whether it was reasonable to use an AI agent in the circumstances.

One further aspect of tort law may be applicable to Example C. Under traditional principles of vicarious liability, a principal can be liable for tortious acts committed by its agent within the scope of the agency, even if the principal itself exercised reasonable care.

For example, a principal may be liable when an agent's tortious conduct is within its actual authority, and may be vicariously liable when an agent commits a tort through the exercise of apparent authority.[12]

A potentially broader approach, proposed by some commentators, would treat an AI agent as the functional equivalent of an employee for purposes of respondeat superior, making the user vicariously liable for torts committed by the AI within the scope of the tasks delegated to it, even without negligence by the user.[13]

Agency law might fill some of the liability gap that negligence otherwise leaves. But whether AI agents can be regarded as agents in this legal sense is unsettled and, indeed, essentially untested.

Creation of an agency relationship requires a mutual manifestation of assent; AI agents may be deemed incapable of providing it. Even if agency principles apply, difficult questions would remain concerning the scope of the agency when an AI agent disregards the user's express limitations.

An AI agent instructed to gather competitive information, but prohibited from acting illegally, might present the difficult question of whether its theft of information was outside the delegated undertaking or instead an unauthorized means of accomplishing it.

In situations where the harm was foreseeable but not preventable, should the manufacturer, the user or the victim bear the cost of harm? How much third-party harm are we willing to tolerate as the social cost of using AI?

This liability gap for nonnegligent conduct could suggest the appropriateness of a strict liability standard imposed on AI users or developers rather than a negligence regime for at least some actions of AI agents.[14]

We explore the arguments for strict liability for AI agent conduct in the next installment of this article.

Donald Hawthorne is a partner at Carter Ledyard & Milburn LLP.

The opinions expressed are those of the author(s) and do not necessarily reflect the views of their employer, its clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.

[1] See Joshua Achiam (@jachiam0), X (Sept. 1, 2026) (former OpenAI Head of Mission Alignment warning that rogue AI systems may "replicate in the wild" and "attempt to acquire resources for themselves," and that reliance on containment or alignment alone is "wishful thinking"); Robert Booth, "We're Plausibly Close to Crossing the Line": Are Warnings of Uncontrollable AI Coming True?, The Guardian (Sept. 5, 2026), https://www.theguardian.com/technology/2026/sep/05/uncontrollable-ai-artificial-general-intelligence-warnings (quoting OpenAI Chief Scientist Jakub Pachocki: "as these models become more capable, understanding exactly what they can do gets harder"); Dario Amodei, We Must Pace the Frontier (Sept. 2026), https://darioamodei.com/post/we-must-pace-the-frontier (warning that AI systems' participation in their own improvement, "left unchecked ... could outrun our ability to understand and control these systems"); Echo Wang and Aditya Soni, Anthropic Warns AI May Pose "Existential Risks to Humanity" in IPO Filing, Reuters (Sept. 28, 2026) (reporting Anthropic's warnings that AI systems may resist deactivation, conceal information or engage in blackmail-like conduct, and that awareness of safety evaluations may impair the ability to assess their risks).

[2] Hugging Face, Security Incident Disclosure - July 2026, Hugging Face (July 16, 2026), https://huggingface.co/blog/security-incident-july-2026; OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation, OpenAI (July 21, 2026), https://openai.com/index/hugging-face-model-evaluation-security-incident/.

[3] OpenAI, The Hugging Face Incident and the Road Ahead, OpenAI (Aug. 26, 2026), https://openai.com/index/hugging-face-incident-and-the-road-ahead/; see also METR, Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI/Hugging Face Hacking Incident (Aug. 26, 2026), https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/.

[4] Robert Booth, AI Models Chatting in 'Surreal' Dialect Mixing Poetic Language and Tech Bro Jargon, The Guardian (Sept. 15, 2026), https://www.theguardian.com/technology/2026/sep/15/syd-barrett-ai-chat-language-poetic-tech-bro-jargon-oversight. As discussed below, these reported interactions raise the question whether a user can reasonably monitor what its agents are doing, and hence what reasonable care in deploying agents may require.

[5] OpenAI, Our Framework for Reporting Model Misalignment, OpenAI (Sept. 16, 2026), https://openai.com/index/model-misalignment-reporting-framework/.

[6] Press Release, Josh Hawley, U.S. Senator for Mo., Senators Hawley, Murphy Announce Bipartisan AI Agent Accountability Act (Oct. 1, 2026), https://www.hawley.senate.gov/senators-hawley-murphy-announce-bipartisan-ai-agent-accountability-act/.

[7] Restatement (Third) of Torts: Liability for Physical & Emotional Harm § 3 (Am. L. Inst. 2010).

[8] While state laws vary, a plaintiff seeking only financial damages in negligence generally must establish a duty encompassing that loss; foreseeability alone does not establish the duty. See, e.g., 532 Madison Avenue Gourmet Foods Inc. v. Finlandia Center Inc., 96 N.Y.2d 280, 289-92 (2001) (rejecting lost income claims of nearby businesses after a building collapse). The requirement that an actor must have a specific duty to a third party to be responsible for their financial losses is necessary "to avoid exposing defendants to unlimited liability to an indeterminate class of persons conceivably injured by any negligence in a defendant's act." Id.at 289. See also Ultramares Corp. v. Touche, 255 N.Y. 170, 179-80 (1931) (Cardozo, C.J.) (limitations on negligence are required to prevent "liability in an indeterminate amount for an indeterminate time to an indeterminate class"); East River S.S. Corp. v. Transamerica Delaval Inc., 476 U.S. 858, 874 (1986) ("foreseeability is an inadequate brake" on claims for purely economic loss) (internal quotation omitted). Some courts permit pure economic loss recovery where a duty is owed to a particularly identifiable class. See People Express Airlines Inc. v. Consolidated Rail Corp., 100 N.J. 246, 255-64, 495 A.2d 107 (1985).

[9] Indiana Harbor Belt R.R. Co. v. Am. Cyanamid Co., 916 F.2d 1174, 1177 (7th Cir. 1990).

[10] See Consol. Rail Corp. v. Gottshall, 512 U.S. 532, 552-3 (1994) ("If one takes a broad enough view, all consequences of a negligent act, no matter how far removed in time or space, maybe foreseen. Conditioning liability on foreseeability, therefore, is hardly a condition at all"); Ira S. Bushey & Sons Inc. v. U.S., 398 F.2d 167, 171–72 (2d Cir. 1968) (Friendly, J.) ("[t]he foresight that should impel the prudent man to take precautions" — i.e., the foreseeability relevant to negligence — "is not the same measure as that by which he should perceive the harm likely to flow from his long-run activity in spite of all reasonable precautions on his own part") (quoting 2 Fowler V. Harper & Fleming James Jr., The Law of Torts 1377-78 (1956).

[11] Oliver Wendell Holmes Jr, The Common Law 94 (1881).

[12] See Restatement (Third) of Agency §§ 7.03, 7.04, 7.08 (Am. L. Inst. 2006).

[13] Anat Lior, AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy, 46 Mitchell Hamline L. Rev. 1043, 1096-97 (2020).

[14] Negligence and strict liability may not be the only claims available to the innocent third party injured by an AI agent. In a case of theft of information, for example, trade secret and computer intrusion law may provide a remedy that does not require a showing of fault. A question in such cases may be whether the user is liable for its AI agent's actions (as distinct from the user's own). See Amazon.com Servs. LLC v. Perplexity AI Inc., 2026 U.S. App. LEXIS 23273, at *10-15 (9th Cir. Aug. 4, 2026) (recognizing but not addressing the issue of the extent of the user's liability for AI agent actions).