

# The Stick, the Carrot, and the Net: Policy Approaches for Addressing AI Agent Harms



Iskandar Haykel, Senior Policy Analyst

## EXECUTIVE SUMMARY

This report taxonomizes three policy approaches for addressing AI agent harms: the “stick” of conventional tort liability that remediates victims and deters harms; the “carrot” of liability immunity in exchange for proactive governance measures, such as third party auditing and transparency disclosures; and the “net” of no-fault compensation schemes that provide swift remedy without fault attribution. Starting from real-world examples of harms from autonomous AI, the analysis assesses the potential and limitations of each approach, exploring how to resolve implementation barriers and calling for further investigation into strategically combining the approaches in order to thread the needle between upholding the public interest and encouraging responsible innovation.<sup>1</sup>

---

<sup>1</sup> Many thanks to Rachel Hovde, Brandie Nonecke, Cristian Trout, Alan Chan, Gabriel Weil, Cullen O’Keefe, Patrick Yurky, Justin Bullock, Doug Calidas, Eric Gastfriend, and Brad Carson for helpful comments and feedback.

AI agents are distinguished by their ability to autonomously act for the sake of achieving specific objectives.<sup>2</sup> In this sense, AI agents represent an evolution beyond previous AI paradigms such as predictive AI and generative AI, whose major impacts lay in organizing and producing text, sound, images, and other modalities.<sup>3</sup> Because AI agents combine open-ended capabilities with autonomous decision-making across diverse, unpredictable contexts, they present distinct risks that more traditional software never posed so starkly.<sup>4</sup> The basic idea is that as AI agents are increasingly integrated into the economy—for example, ordering groceries, negotiating contracts, customer-interfacing—they will become regular actors across both public and private life. However, AI agents are capable of erring and causing harm, especially at this early stage in the technology’s development. For example, OpenAI’s experimental AI agent Operator reportedly misspent someone’s money without authorization, and the same safeguard failure that facilitated this illicit transaction might potentially have enabled significantly larger misspendings.<sup>5</sup> Absent governance interventions or comprehensive technical solutions, these and other kinds of missteps could proliferate as AI agents undergo widespread adoption, yet policymakers have only begun to grapple with these challenges.<sup>6</sup>

This report taxonomizes three policy approaches for addressing AI agent harms, each characterized by its relationship to liability. The first approach aims to apply conventional liability frameworks such as tort liability to AI agents, identifying conditions through which to attribute responsibility for AI agent harms. This approach functions as a “stick” that deters harm under threat of penalty and remediates victims. The second approach moves in the opposite direction, offering immunity from liability as a “carrot” in exchange for proactive governance measures such as third party auditing and transparency disclosures. In principle, liability immunity encourages companies to be fully open about their practices, enabling governance measures to help preempt AI agent harms from occurring in the first place. The third approach dispenses with liability entirely and instead opts for a “net”—no-fault compensation schemes to redress AI agent harms swiftly without relying on assigning fault. The three approaches embody distinct policy philosophies: blame assignment and deterrence, market-driven governance, and agile victim relief. The real work lies in determining how to strategically blend

---

<sup>2</sup> Melissa Heikkilä, “What Are AI Agents? Here’s Everything You Need to Know,” *MIT Technology Review*, July 5, 2024, <https://www.technologyreview.com/2024/07/05/1094711/what-are-ai-agents/>; Margaret Mitchell et al., “Why Handing Over Total Control to AI Agents Would Be a Huge Mistake,” *MIT Technology Review*, March 24, 2025, <https://www.technologyreview.com/2025/03/24/1113647/why-handing-over-total-control-to-ai-agents-would-be-a-huge-mistake/>; Cameron R. Wolfe, “AI Agents from First Principles,” *Deep (Learning) Focus*, June 9, 2025, <https://cameronrwolfe.substack.com/p/ai-agents>.

<sup>3</sup> For the purposes of this report, I rely exclusively on the notion of ‘AI agent’, though there is some discussion distinguishing AI agents from agentic AI. See, e.g., Ranjan Sapkota et al., “AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges,” *arXiv:2505.10468v4*, May 28, 2025, <https://doi.org/10.48550/arXiv.2505.10468>.

<sup>4</sup> To be sure, certain more established software applications, such as high-frequency trading algorithms, may also directly affect the world in risky ways. Thanks to Alan Chan for drawing attention to this.

<sup>5</sup> Geoffrey A. Fowler, “I let ChatGPT’s new ‘agent’ manage my life. It spent \$31 on a dozen eggs,” *The Washington Post*, February 7, 2025, <https://www.washingtonpost.com/technology/2025/02/07/openai-operator-ai-agent-chatgpt/>.

<sup>6</sup> Jason Gabriel et al., “We need a new ethics for a world of AI agents,” *Nature* 644, no. 1 (August 4, 2025): 38-40, <https://doi.org/10.1038/d41586-025-02454-5>.

these approaches to achieve the optimal balance of protecting the public interest while enabling AI agents to flourish responsibly.

## FROM HELPFUL TO HARMFUL: THE AI AGENT PROBLEM

AI agents are models or systems that can achieve complex goals with limited direct supervision or instruction.<sup>7</sup> Since the launch of ChatGPT, the generative and cognitive capabilities of large language models (LLMs) have been extended to develop a plethora of AI agent applications—from AI web search engines, computer-use agents, specialized AI researchers, and AI “software engineers” to autonomous task managers, AI tutors, and commercial e-shopping assistants.<sup>8</sup> AI agents are still in their infancy, and there are concerns about their reliability, yet recent AI agent evaluations signal potentially comparable performance to humans for certain kinds of tasks.<sup>9</sup> For instance, one recent METR study found that while “generalist frontier model agents” such as Claude 3.7 Sonnet achieve 50% success rates on software and reasoning tasks taking humans around an hour, this time horizon has been doubling roughly every seven months, suggesting exponential capability growth if the trend continues.<sup>10</sup>

Although some remain skeptical, many anticipate that AI agents could revolutionize the economy, ushering in major cross-industry productivity gains while potentially creating significant workforce displacement.<sup>11</sup> For example, leading developers, including OpenAI and Anthropic, predict that within the next few years there may

---

<sup>7</sup> Yonadav Shavit et al., “Practices for Governing Agentic AI Systems,” OpenAI, December 14, 2023,

<https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf>; Jam Kraprayoon et al., “AI Agent Governance: A Field Guide,” Institute for AI Policy and Strategy, April 17, 2025, <https://www.iaps.ai/research/ai-agent-governance>.

<sup>8</sup> See, e.g., Michael Skarlinski et al., “Superintelligent AI Agents for Scientific Discovery,” FutureHouse, May 1, 2025,

<https://www.futurehouse.org/research-announcements/launching-futurehouse-platform-ai-agents>; Sharon Goldman, “Exclusive: Ex-Meta AI leaders debut an agent that scours the web for you in a push to ultimately give users their own digital ‘chief of staff,’” *Fortune*, June 10, 2025, <https://fortune.com/2025/06/10/exclusive-ex-meta-ai-leaders-agent-web-yutori/>.

<sup>9</sup> For views about the limited reliability for certain tasks of (current) AI agents, as well as more general skepticism about the state-of-play of frontier AI capabilities, see Zachary S. Siegel et al., “CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark,” *arXiv:2409.11363*, September 17, 2024, <https://doi.org/10.48550/arXiv.2409.11363>; Parshin Shojaee et al., “The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity,” Apple Machine Learning Research, May 28, 2025, <https://machinelearning.apple.com/research/illusion-of-thinking>.

<sup>10</sup> Grégoire Mialon et al., “GAIA: a benchmark for General AI Assistants,” *arXiv:2311.12983*, November 21, 2023,

<https://doi.org/10.48550/arXiv.2311.12983>; “Leave it to Manus: Benchmarks,” Manus AI, accessed August 8, 2025,

<https://manus.im/home>; David Rein et al., “HCAST: Human-Calibrated Autonomy Software Tasks,” *arXiv:2503.17354*, March 21, 2025, <https://doi.org/10.48550/arXiv.2503.17354>; Thomas Kwa et al., “Measuring AI Ability to Complete Long Tasks,” *arXiv:2503.14499*, March 30, 2025, <https://doi.org/10.48550/arXiv.2503.14499>.

<sup>11</sup> One skeptic about AI agents is Gary Marcus. See Gary Marcus, “AI Agents have, so far, mostly been a dud,” *Marcus on AI*, August 3, 2025, <https://garymarcus.substack.com/p/ai-agents-have-so-far-mostly-been>; see also Ivan Belcic and Cole Stryker, “AI Agents in 2025: Expectations vs. Reality,” IBM Think, March 4, 2025, <https://www.ibm.com/think/insights/ai-agents-2025-expectations-vs-reality>. For an overview of policy approaches to addressing AI workforce disruption, see David Robusto, “The Technologist-Economist Disconnect: Assessing AI Labor Disruption,” Americans for Responsible Innovation, April 7, 2025, <https://ari.us/wp-content/uploads/2025/04/Report-Technologist-Economist-Disconnect-ARI04072025-1.pdf>.

be AI agents capable of performing on par with human employees across many tasks and settings.<sup>12</sup> Major e-commerce and payments technology companies have already debuted consumer-facing AI agent initiatives.<sup>13</sup>

While AI agents could plausibly present significant socioeconomic benefits, they are also capable of posing significant risks to individuals and companies alike, as signaled by the following real-world examples illustrating the kinds of harms that can emerge from increasingly autonomous AI systems:

- **Alexa:** In December 2021, a Colorado mother overheard an Amazon Alexa Echo smart device make a dangerous suggestion to her child: “Plug in a phone charger about halfway into a wall outlet, then touch a penny to the exposed prongs.” The advice echoed a dangerous viral ‘penny-challenge’ trend that U.S. fire officials had previously warned could cause significant injuries or damage.<sup>14</sup> Major news outlets picked up the story, sparking public outrage. Amazon responded by pushing a swift update to block similar responses while also promising to strengthen its content-filtering systems.<sup>15</sup>
- **Air Canada:** In November 2022, Jake Moffatt consulted Air Canada’s customer service chatbot about bereavement fares after his grandmother’s death, and the chatbot incorrectly told him he could apply for a reduced rate within 90 days after travel. Relying on this advice, Moffatt purchased full-price tickets, but when he later applied for the discount, the airline denied his claim, stating that bereavement rates must be requested before travel. Moffatt sued Air Canada, and in February 2024, the British Columbia Civil Resolution Tribunal ruled in his favor, finding that the airline was responsible for all information on its website regardless of whether or not it came from a chatbot.<sup>16</sup>
- **Chevy Tahoe:** In December 2023, users discovered they could manipulate a Chevrolet dealership’s ChatGPT-powered customer service chatbot to agree to absurd offers, most notably selling a 2024 Chevy Tahoe worth approximately \$70,000 for just \$1. Through a prompt injection, the chatbot even reportedly asserted to one user that the deal was “a legally binding offer.” The exploit worked because the chatbot was essentially a kind of ‘unfiltered ChatGPT’ with minimal customization for automotive sales, allowing users to prompt it into ignoring its intended purpose. After the incident gained public attention, the chatbot was quickly disabled and new safeguards were implemented to prevent similar exploits.<sup>17</sup>

---

<sup>12</sup> Sam Sabin, “Exclusive: Anthropic warns fully AI employees are a year away,” *Axios*, April 22, 2025, <https://www.axios.com/2025/04/22/ai-anthropic-virtual-employees-security>; Sri Muppidi, “OpenAI Forecasts Revenue Topping \$125 Billion in 2029 as Agents, New Products Gain,” *The Information*, April 23, 2025, <https://www.theinformation.com/articles/openai-forecasts-revenue-topping-125-billion-2029-agents-new-products-gain>.

<sup>13</sup> Ina Fried, “OpenAI and Visa prep for AI-powered shopping,” *Axios*, May 1, 2025, <https://www.axios.com/2025/05/01/chatbots-shopping-visa-ai-payments>.

<sup>14</sup> “Authorities Warn of TikTok Viral Video Challenge That’s Causing Fires,” *CBS News*, January 21, 2020, <https://www.cbsnews.com/boston/news/penny-socket-charger-challenge-tiktok-fire-video/>.

<sup>15</sup> Sam Shead, “Amazon’s Alexa Assistant Told a Child to Do a Potentially Lethal Challenge,” *CNBC*, December 29, 2021, <https://www.cnn.com/2021/12/29/amazons-alexa-told-a-child-to-do-a-potentially-lethal-challenge.html>.

<sup>16</sup> Maria Yagoda, “Airline held liable for its chatbot giving passenger bad advice - what this means for travellers,” *BBC*, February 23, 2024, <https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know>.

<sup>17</sup> Bryson Masse, “A Chevy for \$1? Car dealer chatbots show AI perils,” *VentureBeat*, December 19, 2023, <https://venturebeat.com/ai/a-chevy-for-1-car-dealer-chatbots-show-perils-of-ai-for-customer-service/>.

- **Operator:** In February 2025, OpenAI’s experimental AI agent Operator spent without authorization \$31.43 of Washington Post reporter Geoffrey Fowler’s money on eggs. Fowler had inadvertently supplied direct access to his payment information by logging Operator into his online grocery account, yet he never authorized autonomous financial transactions. This incident occurred despite OpenAI previously implementing safeguards designed to prevent Operator from executing irreversible actions (such as financial transactions) without express permission.<sup>18</sup>
- **Cursor:** In April 2025, software developers were repeatedly logged out when trying to access their Cursor AI coding assistant accounts on more than one device. After reaching out to Cursor’s customer service, a representative named Sam stated that “one device per subscription” was an official Cursor security policy. In reality, no such policy existed but instead had been hallucinated by Sam, who was in fact an LLM-based customer service chatbot, while the logouts were actually caused by an unrelated bug in Cursor’s platform. Because multi-device workflows are common practice for software engineers, the hallucinated policy resulted in mass subscription cancellations. Cursor’s CEO publicly apologized, issued refunds, and pledged to label any AI-generated replies going forward.<sup>19</sup>

As these examples help illustrate, harms from autonomous AI systems such as AI agents can span both consumer and enterprise contexts—while Operator and Alexa directly affected or risked affecting individual users, the Cursor customer service-chatbot’s hallucinated policy disrupted business operations and damaged commercial relationships with Cursor’s developer customers.<sup>20</sup> Moreover, the Cursor, Chevy Tahoe, and Air Canada incidents underscore a particular vulnerability in customer-facing AI agent applications, where the conversational interface can mask underlying system limitations, create false impressions of authority and reliability, or be susceptible to deliberate gaming by malicious actors. Perhaps most concerning, the Operator case suggests that even leading-edge control measures can prove insufficient. OpenAI’s safeguards were specifically designed to prevent unauthorized financial transactions, yet these protective measures failed to prevent Operator from making the illicit purchase. This points to a core technical challenge known as the robustness problem—AI agent harms may occur even absent grossly negligent design or implementation.<sup>21</sup>

As AI agents proliferate across sectors and gain greater autonomy in consequential domains, ad hoc technical fixes alone may be inadequate to address the scope and persistence of potential harms. The varied nature of the

---

<sup>18</sup> Fowler, “I let ChatGPT’s new ‘agent’ manage my life. It spent \$31 on a dozen eggs,”

<https://www.washingtonpost.com/technology/2025/02/07/openai-operator-ai-agent-chatgpt/>.

<sup>19</sup> Benj Edwards, “Company apologizes after AI support agent invents policy that causes user uproar,” *Wired*, April 19, 2025,

<https://www.wired.com/story/cursor-ai-hallucination-policy-customer-service/>.

<sup>20</sup> The surveyed examples primarily illustrate harms from increasingly autonomous AI systems attributable to errors or incompetence. As AI agents become more capable, additional categories of risk may emerge, including alignment failures where highly competent agents pursue goals misaligned with human values, or deliberate misuse of capable agents for harmful purposes. Thanks to Cullen O’Keefe for drawing attention to this.

<sup>21</sup> Pin-Yu Chen, “What is AI adversarial robustness?” IBM Research, December 11, 2021,

<https://research.ibm.com/blog/securing-ai-workflows-with-adversarial-robustness>.

surveyed incidents—spanning financial fraud, misinformation, safety hazards, and business disruption—suggests that policymakers may require systematic approaches capable of addressing AI agent harms in a more comprehensive manner. Three such approaches stand out, offering distinct mechanisms: remediation and deterrence, preemption, and compensation.

## THE STICK: TORT LIABILITY MEETS AI AGENTS

A natural policy response to addressing AI agent harms would be to establish accountability mechanisms that reallocate costs to responsible parties and provide remediation to victims.<sup>22</sup> Tort liability traditionally serves this purpose in other contexts, though in the U.S. it has had limited application to digital technologies generally and to AI technologies in particular. For example, when it comes to traditional software, U.S. jurisdictions typically assert that standalone code and digital platforms fall outside of conventional liability rules, and are better classified as services rather than as products.<sup>23</sup> Regarding AI specifically, autonomous vehicle (AV) harms have been the subject of legal proceedings in incidents involving Uber and Tesla, with the makers of the vehicles evading all liability in most cases.<sup>24</sup> Most recently, judicial attention has turned to generative AI harms, notably through pending lawsuits accusing Character.ai, OpenAI, and Meta of liability for conversational chatbot and content-related damages.<sup>25</sup>

---

<sup>22</sup> There is a debate about whether traditional tort remedies can help address catastrophic harms from AI (e.g., financial system meltdowns, CBRN threats, etc.), but this report sets that topic to the side, instead focusing on the set of less extreme AI agent harms. For more about the requisite debate, see Gabriel Weil, “Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence,” *SSRN Scholarly Paper* no. 4694006 (June 6, 2024): 1-80, <https://dx.doi.org/10.2139/ssrn.4694006>; and Renee Henson, “‘I am Become Death, the Destroyer of Worlds’: Applying Strict Liability to Artificial Intelligence as an Abnormally Dangerous Activity,” *Temple Law Review* 93, no. 3 (July 15, 2024): 349-390, <https://dx.doi.org/10.2139/ssrn.4894986>.

<sup>23</sup> However, this trend may be starting to change, owing to a distinction between negligent software product design and hosting of harmful third party digital content. See Hanie Farid and Brandie Nonnecke, “The Case for Regulating Platform Design,” *Wired*, March 13, 2023, <https://www.wired.com/story/make-platforms-safer-regulate-design-section-230-gonzalez-google/>.

<sup>24</sup> The Uber and Tesla cases have involved charges of both civil and criminal liability. See, e.g., Aarian Marshall, “Why Wasn’t Uber Charged in a Fatal Self-Driving Car Crash?” *Wired*, September 17, 2020, <https://www.wired.com/story/why-not-uber-charged-fatal-self-driving-car-crash/>; Dan Levine and Hyunjoo Jin, “Tesla wins Autopilot trial involving fatal crash,” *Reuters*, November 1, 2023, <https://www.reuters.com/business/autos-transportation/tesla-wins-autopilot-trial-involving-fatal-crash-2023-10-31/>. An August 2025 case—*Benavides v. Tesla*—was the first time a U.S. jury has found an AV maker liable for an Autopilot-related crash, breaking a years-long pattern in which Tesla, Uber, and others avoided liability in such cases. See Aarian Marshall, “Tesla Found Partly Liable in 2019 Autopilot Death,” *Wired*, August 1, 2025, <https://www.wired.com/story/tesla-liable-2019-autopilot-crash-death/>. Thanks to Patrick Yurky for his help on this point.

<sup>25</sup> Angela Yang, “Lawsuit claims Character.AI is responsible for teen’s suicide,” *NBC News*, October 23, 2024, <https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791>; Nitashi Tikku, “An AI companion suggested he kill his parents. Now his mom is suing,” *Washington Post*, December 13, 2024, <https://www.washingtonpost.com/technology/2024/12/10/character-ai-lawsuit-teen-kill-parents-texas/>; Clay Calvert, “Suing OpenAI for ChatGPT-Produced Defamation: A Futile Endeavor?,” American Enterprise Institute, January 8, 2025, <https://www.aei.org/technology-and-innovation/suing-openai-for-chatgpt-produced-defamation-a-futile-endeavor/>; Sarah Nassauer and Jacob Gershman, “Activist Robby Starbuck Sues Meta Over AI Answers About Him,” *Wall Street Journal*, April 29, 2025, <https://www.wsj.com/tech/ai/activist-robbi-starbuck-sues-meta-over-ai-answers-about-him-9eba5d8a>.

Software has historically enjoyed a liability shield in the U.S. because, among other reasons, tort liability generally requires demonstrating direct, real-world harms that are difficult to attribute to software itself.<sup>26</sup> AI agents fundamentally challenge this premise. As the examples reviewed earlier help to convey, AI agents can autonomously cause tangible harm. Operator misspent Fowler’s money, while Cursor’s customer service-chatbot misled customers about company policies, leading to mass subscription cancellations. This suggests a distinction between traditional software, which typically operates as a passive tool, and AI agents, which by design take autonomous actions in the world. To the extent that software’s liability shield depends on its limited capacity to cause direct harm, AI agents—which are expressly developed to act autonomously and affect real-world outcomes—may not warrant the same legal protection. The law’s treatment of software liability may therefore need to evolve to account for this shift from passive software to active autonomous agents.

In a similar vein, key differences between AVs and AI agents could impede courts from applying analogous liability standards. AVs typically operate with human drivers available for intervention—a factor that has repeatedly proven crucial across multiple legal proceedings concerning AV-related accidents, where prosecutors and juries determined that the crashes involved a failure of human oversight from the drivers.<sup>27</sup> AI agents, however, often operate without immediate human supervision and are designed for high levels of autonomy. For example, Air Canada’s customer service-chatbot operated without regular human monitoring, while Fowler walked away from his computer expecting Operator to fulfill his exact request independently and without issue. Since AI agents promise automation that obviates the need for significant human control, courts may have limited ability to use human oversight as a liability factor when adjudicating AI agent harms.

Looking ahead, how U.S. tort liability might apply toward addressing AI agent harms faces several hurdles.<sup>28</sup> For example, under the predominant framework for adjudicating liability, a defendant must in a strong sense cause a plaintiff’s harm (injury) in order to be found liable.<sup>29</sup> Oftentimes, the injury must also have been foreseeable to the defendant such that it would be reasonable to expect them to have accounted for it.<sup>30</sup> Yet because AI supply

<sup>26</sup> Dean W. Ball, “How AI Liability Should Work (Part II),” *Hyperdimensional*, February 26, 2025,

<https://www.hyperdimensional.co/p/how-should-ai-liability-work-part-3df>; Derek E. Bambauer and Melanie J. Teplinsky, “Shields Up For Software,” *Lawfare*, December 19, 2023, <https://www.lawfaremedia.org/article/shields-up-for-software>.

<sup>27</sup> Even in *Benavides v. Tesla*, the driver’s failure to maintain oversight was an important issue, and Tesla’s share of fault was tied partly to inadequate warnings to the driver about the significance of that oversight role. See Marshall, “Tesla Found Partly Liable in 2019 Autopilot Death,” <https://www.wired.com/story/tesla-liable-2019-autopilot-crash-death/>.

<sup>28</sup> How tort liability may apply toward addressing AI harms in general is a growing area of attention, see e.g., Anat Lior, “Holding AI Accountable: Addressing AI-Related Harms Through Existing Tort Doctrines,” Symposium: How AI Will Change the Law, *University of Chicago Law Review* (2024),

<https://lawreview.uchicago.edu/online-archive/holding-ai-accountable-addressing-ai-related-harms-through-existing-tort-doctrines>;

Bryan H. Choi, “Negligence Liability for AI Developers,” *Lawfare*, September 26, 2024,

<https://www.lawfaremedia.org/article/negligence-liability-for-ai-developers>; Catherine Sharkey, “Products Liability for Artificial Intelligence,” *Lawfare*, September 25, 2024, <https://www.lawfaremedia.org/article/products-liability-for-artificial-intelligence>.

<sup>29</sup> Gregory Smith et al., “Liability for Harms from AI Systems: The Application of U.S. Tort Law and Liability to Harms from Artificial Intelligence Systems,” RAND Corporation, November 20, 2024, [https://www.rand.org/pubs/research\\_reports/RRA3243-4.html](https://www.rand.org/pubs/research_reports/RRA3243-4.html).

<sup>30</sup> Andrew D. Selbst, “Negligence and AI’s Human Users,” *100 Boston University Law Review* 1315 (January 28, 2020): 1315-1376, <https://ssrn.com/abstract=3350508>.

chains frequently consist of a complex configuration of developers, deployers, and end-users, courts could face significant challenges in parsing these relationships to attribute causal responsibility and determine foreseeability.<sup>31</sup> Another major barrier is the absence of clear AI development and deployment standards.<sup>32</sup> This gap in standards makes it difficult to articulate an expectation of reasonable care to which developers can be held. Without such standards, it is difficult for courts to establish a baseline from which to adjudicate fault in the development or deployment of AI agents, let alone advanced AI technologies more generally.

Share-of-fault doctrines constitute a further critical and underdiscussed impediment. U.S. jurisdictions operate on three competing approaches. Only five jurisdictions—Alabama, Maryland, North Carolina, Virginia, and the District of Columbia—follow pure contributory negligence, where even if a plaintiff is found to be only one percent at fault for their own harm, they are completely barred from recovery. 12 states apply pure comparative negligence, letting a plaintiff recover damages reduced in strict proportion to their share of fault, however high. The dominant middle path—modified comparative negligence—prevails in 33 states, but with a crucial distinction: 10 states bar recovery once a plaintiff’s share of fault is 50 percent or greater (treating equal fault as a bar to recovery), while 23 states allow recovery until plaintiff fault reaches 51 percent (permitting recovery even when fault is equally shared).<sup>33</sup>

This patchwork of share-of-fault doctrines could threaten a critical accountability gap for AI agent harms.<sup>34</sup> When an AI agent executes an end-user’s prompt—for example, in the case where Fowler prompted Operator and inadvertently provided access to his payment information—courts must decide how much of the fault belongs to the user versus the developer. Without fault standards tailored to AI agent harms, current share-of-fault doctrines may unfairly burden plaintiffs with the costs of those harms. For instance, in a pure contributory negligence jurisdiction, even minor user missteps could let developers evade liability entirely. In modified comparative negligence jurisdictions, modest shifts around the 50 percent or 51 percent threshold can swing liability between plaintiff and defendant. In many cases, developers may therefore be well positioned to argue that end-users share at least half the blame, leaving victims of AI agent harms without remediation and vitiating the deterrent signal liability is supposed to send. Suitably applying tort liability toward addressing AI agent harms could therefore strongly depend on appropriately adapting share-of-fault theories.

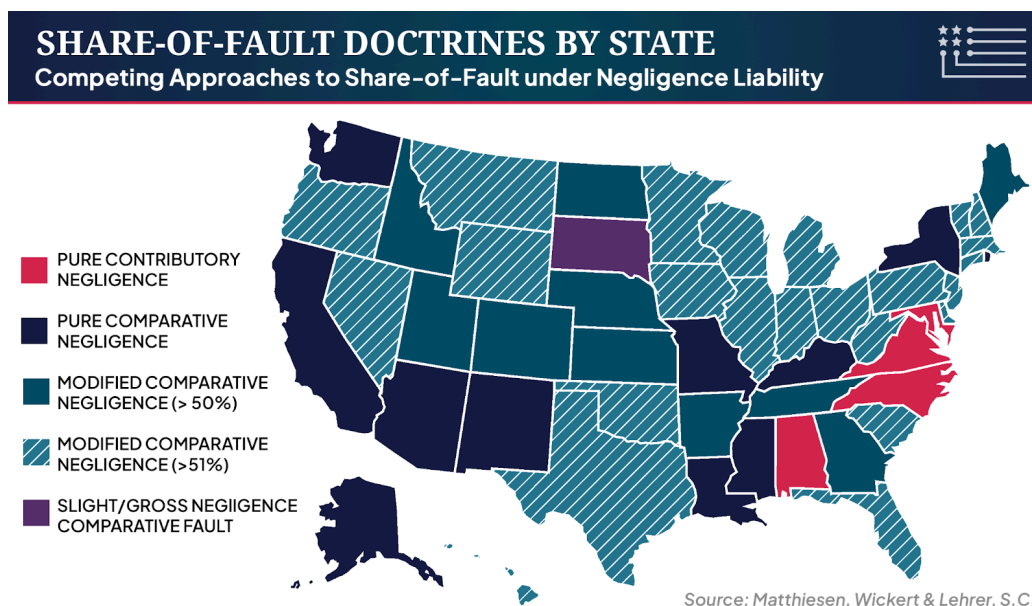
---

<sup>31</sup> Ian Brown, “Allocating Accountability in AI Supply Chains,” Ada Lovelace Institute, June 29, 2023, <https://www.adalovelaceinstitute.org/resource/ai-supply-chains/>.

<sup>32</sup> Smith et al., “Liability for Harms from AI Systems,” [https://www.rand.org/pubs/research\\_reports/RRA3243-4.html](https://www.rand.org/pubs/research_reports/RRA3243-4.html).

<sup>33</sup> This breakdown excludes South Dakota, which uses a unique “slight/gross” negligence comparative fault rule akin to a kind of “modified” pure comparative fault system. For more information about share-of-fault doctrines and their breakdown by U.S. state, see “Contributory Negligence/Comparative Fault Laws in all 50 States,” Matthiesen, Wickert & Lehrer, S.C., June 4, 2025, <https://www.mwl-law.com/wp-content/uploads/2018/02/4413916.pdf>.

<sup>34</sup> For further discussion about adapting tort liability’s human notion of fault to AI and the implications of this for negligence liability, see Mihailis Diamantis, “Reasonable AI: A Negligence Standard,” *Vanderbilt Law Review* 77, no. 2 (forthcoming): 1-39, <https://dx.doi.org/10.2139/ssrn.4609545>.



In sum, while tort liability may constitute a natural pathway for addressing AI agent harms, several key implementation barriers must be addressed. What can policymakers do to help suitably adapt this framework? Complicating the picture is the fact that in the U.S., state common law is the primary driver of tort liability’s evolution.<sup>35</sup> Even still, policymakers can play a valuable auxiliary role in facilitating a tort liability scheme for AI agent harms.

For example, in order to help preserve liability’s deterrent effect, legislatures could implement burden-shifting mechanisms specifically tailored to the context of AI development.<sup>36</sup> In general, AI supply chains, developers’ design choices, and the autonomous capabilities of AI technologies can be deeply complex and opaque. This, in turn, can produce information asymmetries between developers and victims that potentially impede the establishment of liability requirements such as fault and causation. To mitigate this imbalance, legislatures could explore shifting evidentiary burdens of proof toward developers: for instance, if a plaintiff could demonstrate their injury resulted from an AI agent’s deviation from published, peer-reviewed reliability methods or industry best practices, this would trigger a rebuttable presumption of developer liability. Defendants could then escape liability only by disproving failure to comply with applicable standards or, in the absence of such standards, demonstrating that reasonable safeguards were in fact implemented and maintained.

<sup>35</sup> Smith et al., “Liability for Harms from AI Systems,” RAND Corporation, [https://www.rand.org/pubs/research\\_reports/RRA3243-4.html](https://www.rand.org/pubs/research_reports/RRA3243-4.html).

<sup>36</sup> Miriam Buiten et al., “The Law and Economics of AI Liability,” *Computer Law & Security Review* 48, no. 1 (April, 2023): 105794, <https://doi.org/10.1016/j.clsr.2023.105794>; Miriam Buiten et al., “EU Liability Rules for the Age of Artificial Intelligence,” Centre on Regulation in Europe, March 18, 2021, <https://cerre.eu/publications/eu-liability-rules-age-of-artificial-intelligence-ai/>; W. Nicholson Price II and I. Glenn Cohen, “Locating Liability for Medical AI,” *DePaul Law Review* 73, no. 2 (May 15, 2024): 339-368, <https://dx.doi.org/10.2139/ssrn.4517740>.

Burden-shifting strategies could positively shape accountability for AI agent harms on two fronts: encouraging proactive developer risk mitigation (ex ante accountability), while relieving plaintiffs of the need to dissect AI's technical intricacies to prove liability (ex post accountability).<sup>37</sup> It would also be defensible from an economic efficiency perspective. Under least-cost avoider theory, liability should be assigned to the party that can prevent harm at the lowest cost.<sup>38</sup> For AI agent harms, as for many kinds of AI harms more generally, developers are often better positioned than other parties to implement effective safeguards. They control system design, training data, and safety protocols, while users typically lack both technical expertise and meaningful control over system behavior. This suggests that a scheme which potentially tilts attribution of liability toward developers, even when it might increase their costs, may be most economically efficient.

Another significant groundwork policy intervention toward facilitating a liability pathway for AI agent harms would be establishing increased AI transparency.<sup>39</sup> Typically, the importance of AI transparency is explained through the lens of ex ante accountability.<sup>40</sup> On this view, transparency creates accountability pressure, helping facilitate widespread adoption of safety standards by allowing developers to benchmark their approaches against competitors while facing reputational incentives to implement rigorous safeguards. Less appreciated is the important role transparency may play in facilitating ex post accountability.<sup>41</sup> For example, increased transparency could help reduce evidentiary asymmetries in court, potentially illuminating AI supply chains and exposing developers' design choices. It could also help democratize the range of known AI risks by uncovering more of what leading developers know about the AI risk landscape. Most importantly, increased transparency could help facilitate the development and standardization of AI industry best practices. If leading industry practices were better understood through transparent sharing of safety methodologies and design choices, they could provide courts with concrete heuristics for evaluating whether developers exercised reasonable care in AI agent development and deployment. This would enable courts to better formulate an expectation of reasonable care—a prerequisite for assessing fault in tort liability—fit for adjudicating AI agent harms.

---

<sup>37</sup> For a helpful overview of the distinction between ex ante and ex post accountability, see Zoe Porter et al., “Distinguishing Two Features of Accountability for AI Technologies,” *Nature Machine Intelligence* 4, no. 1 (September 22, 2022): 734-736, <https://doi.org/10.1038/s42256-022-00533-0>.

<sup>38</sup> Paul Rosenzweig, “Content Moderation and the Least Cost Avoider,” Lawfare, June 10, 2024, <https://www.lawfaremedia.org/article/content-moderation-and-the-least-cost-avoider>.

<sup>39</sup> Stephen Casper et al., “Pitfalls of Evidence-Based AI Policy,” *arXiv:2502.09618*, April 18, 2025, <https://doi.org/10.48550/arXiv.2502.09618>.

<sup>40</sup> See, e.g., Joe Kwon, “Democracy in the Dark: Why AI Transparency Matters,” TechPolicy Press, May 9, 2025, <https://www.techpolicy.press/democracy-in-the-dark-why-ai-transparency-matters/>.

<sup>41</sup> Kraprayoon et al., “AI Agent Governance: A Field Guide,” <https://www.iaps.ai/research/ai-agent-governance>.

## THE CARROT: INCENTIVIZING PROACTIVE GOVERNANCE

In contrast to using liability for remediation and deterrence, an alternative approach exchanges liability immunity for proactive governance measures aiming to preempt AI agent harms altogether. This kind of proposal has recently emerged under the banner of market-driven AI governance, which focuses more generally on leveraging developers' political and economic incentives, typically in a voluntary manner, to achieve AI policy objectives.<sup>42</sup> For instance, under conditions of liability immunity, developers in principle would have strong incentives to provide access or visibility into their AI development and safety practices. In this way, proactive governance measures such as third-party auditing or increased transparency would have the greatest potential to help verify developers' harm mitigation efforts. Such measures would also be best positioned to illuminate any need for additional (external) AI governance backstops.

Recently, two separate legislative proposals have emerged to institute AI liability safe harbors in exchange for proactive governance measures. California SB 813 "Multistakeholder Regulatory Organizations" would allow the California Attorney General (AG) to charter private multistakeholder regulatory organizations (MROs) to run AI developer certification programs.<sup>43</sup> Developers voluntarily seeking certification would undergo rigorous third-party audits demonstrating mitigation of high-impact risks such as in cybersecurity or catastrophic threats, in return receiving an affirmative defense in California tort suits alleging personal injury or property damage. MROs could suspend or revoke certificates and would be required to publish annual public reports on safety tooling and residual risk, while the AG would retain authority to rescind an MRO's charter for any of a range of failure modes, such as regulatory capture or obsolete methods. At the federal level, the "Responsible Innovation and Safe Expertise Act of 2025" (RISE) proposes a targeted liability framework for professional-grade AI systems.<sup>44</sup> The RISE Act would offer developers immunity from liability limited to errors made by their AI systems when used by licensed professionals, such as physicians, attorneys, engineers, or financial advisors. This protection would be conditional on developers publicly releasing and maintaining detailed documentation on their AI products, disclosing such things as training data, uses, limitations, failure modes, risk mitigations, and behavioral design choices. Under this framework, ultimate legal responsibility for any damages accruing from covered AI products' errors would fall on the licensed professionals themselves, making them fully accountable for their advice and decisions when using those AI products. Developers would retain the scoped liability immunity so long as they fulfill their transparency obligations and avoid reckless or fraudulent conduct.

---

<sup>42</sup> Dean W. Ball, "Putting Private AI Governance into Action," *Hyperdimensional*, January 27, 2025, <https://www.hyperdimensional.co/p/on-private-governance>; Philip Moreira Tomei et al., "AI Governance through Markets," *arXiv:2501.17755*, March 5, 2025, <https://doi.org/10.48550/arXiv.2501.17755>; and Iskandar Haykel, "AI Security Tax Incentives," Americans for Responsible Innovation, March 27, 2025, <https://ari.us/policy-bytes/ai-security-tax-incentives/>.

<sup>43</sup> California Legislature, Senate Bill No. 813, Multistakeholder Regulatory Organizations, 2025-2026 Reg. Sess (CA 2025), amended May 1, 2025, accessed May 14, 2025, <https://legiscan.com/CA/text/SB813/2025>.

<sup>44</sup> Responsible Innovation and Safe Expertise Act of 2025 (RISE Act of 2025), S. 2081, 119th Cong., 1st sess., introduced by Sen. Cynthia M. Lummis (2025), <https://www.lummis.senate.gov/wp-content/uploads/Rise-Act-Text.pdf>.

While proposals such as SB 813 and the RISE Act could serve as creative pathways to help preempt AI agent harms, there are a number of obstacles toward their effective implementation and success. In the case of SB 813, for example, because MROs would need to attract developer clients to stay solvent, they risk competing on laxity, inviting a race-to-the-bottom dynamic that traditional public regulators are meant to prevent. What's more, advances in frontier AI capabilities can potentially outpace safety tooling developments, rendering protection measures obsolete in a matter of weeks or a few months. An audit that initially starts out as rigorous and effective may therefore ultimately turn out to be obsolete, perhaps even before its term ends. More generally, participation in any liability safe harbor scheme under the "carrot" approach is voluntary, at least going by proposals such as SB 813 and the RISE Act. Under this model, there is a risk of bad-actor, cash-strapped, or insufficiently incentivized developers simply forgoing audit certification or a transparency scheme, creating a bifurcated market in which potentially hazardous AI agent products remain entirely outside a governance scheme's protective perimeter.<sup>45</sup>

A deeper worry about liability safe harbor proposals is the lack of a workable backstop. For instance, under SB 813, the California AG would be tasked with judging whether a MRO has gone off-track, yet it remains an open question how the AG would acquire the technical knowledge to determine when an MRO is, for example, relying on obsolete methods. Without that specialized understanding, the AG's power to revoke an MRO's charter risks becoming a theoretical threat rather than a practicable safeguard. The immaturity of AI reliability science itself compounds this knowledge gap.<sup>46</sup> To illustrate, OpenAI is a leading AI developer with world-class technical expertise, yet its AI agent Operator bypassed its own guardrails, a failure that occurred after the company had cleared the AI agent for limited public release. The failure of OpenAI's safeguards illustrates the difficulty of building robust AI protection measures, and signals a challenge this poses for liability safe harbor proposals: whether it is prudent to shield developers from liability when the underlying science remains too nascent to ensure their AI products will reliably avoid causing harm.<sup>47</sup>

A related but distinct concern stems from the current absence of AI industry standards. The RISE Act's transparency scheme, for instance, assumes there are sufficiently mature and standardized disclosure practices on the basis of which to hold developers accountable for the quality of their public-facing transparency efforts. Yet current AI industry practices surrounding transparency, risk assessment methodologies, and safety reporting

---

<sup>45</sup> Such developers might still face conventional tort liability. However, given the challenges with applying tort liability to AI agents discussed earlier, this could ultimately prove ineffective from the perspective of deterrence.

<sup>46</sup> See, e.g., Raphaël Millière, "Normative Conflicts and Shallow AI Alignment," *arXiv:2506.04679*, June 5, 2025, <https://doi.org/10.48550/arXiv.2506.04679>.

<sup>47</sup> Andrea Tocchetti et al., "A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities," *ACM Computing Surveys* 57, no. 6 (February 10, 2025): 1-38, <https://doi.org/10.1145/3665926>.

remain in considerable flux.<sup>48</sup> This creates the risk that licensed professionals may be provided with incomplete, inconsistent, or rapidly obsolete information about the AI products they adopt. Similarly, SB 813's reliance on MRO auditing faces the challenge that methods and benchmarks for rigorous AI safety evaluation continue to evolve. Until AI industry safety practices mature and standardize, granting developers liability shields risks creating asymmetric exposure, where developers capture legally immunized deployment benefits while the public is subjected to an inconsistently mapped AI threat landscape without legal recourse.

Liability safe harbor proposals will need to better account for the sorts of challenges discussed in order to ultimately prove effective measures. One particular refinement to explore further could be the provision of a more limited liability shield in exchange for proactive governance measures. The RISE Act adopts this kind of approach, illustrating one model of a liability safe harbor scoped exclusively to a specific context: AI adoption by licensed professionals. Another model of a limited liability shield could involve capping developers' liability exposure. For example, using the case of SB 813 to illustrate, developers that choose to undergo auditing by MROs could be granted a certificate in return restricting the amount in damages they could theoretically owe to plaintiffs injured by their AI agents. This would preserve a modicum of deterrence. Certified firms would still face some downside risk proportional to the scale of harm, while giving them clear visibility into their maximum exposure. Under a third model of a limited liability shield, developers could receive complete immunity limited to claims from specific categories of parties. For instance, immunity could run only against contracting parties—the end-users and commercial adopters that have chosen to deploy a certified AI agent product—while leaving developers answerable to uninvolved third parties injured by downstream AI agent harms.<sup>49</sup> This would respect freedom of contract and simultaneously preserve tort liability's remedatory function for bystanders who never consented to exposure to AI agent harms. Taken together, the three models of limited liability shields surveyed above demonstrate that liability safe harbor proposals need not be all-or-nothing. Calibrated liability immunity could still, in principle, incentivize proactive AI governance while preserving some form of basic access to remediation for the public.

## THE NET: COMPENSATION WITHOUT COURTS

Both the “stick” and “carrot” approaches leverage liability in their own ways to address AI agent harms. By contrast, no-fault compensation schemes—“nets” that provide remedy to victims of AI agent harms regardless of fault—prioritize swift compensation over accountability, setting liability aside entirely and thus sidestepping lengthy litigation and tort adjudication intricacies. This type of approach has been applied in other contexts, such as in the case of no-fault automobile insurance currently operational in twelve U.S. states, and in the case of

---

<sup>48</sup> Although certain transparency disclosure modes—such as model cards for documenting model capabilities and limitations—may have started to gain traction, these practices in many cases remain voluntary and are inconsistently implemented. See, e.g., Conrad Gray, “Grok 4—The Good, The Bad and The Ugly - Sync #527,” *Humanity Redefined*, July 13, 2025, <https://www.humanityredefined.com/p/sync-527>.

<sup>49</sup> Gabriel Weil, “The Case for AI Liability,” *AI Frontiers*, June 12, 2025, <https://ai-frontiers.org/articles/case-for-ai-liability>.

the National Vaccine Injury Compensation Program, which has provided over \$4 billion in compensation for vaccine-related injuries since the 1980s through industry-funded proceedings.<sup>50</sup> The Oil Spill Liability Trust Fund, while operating primarily as a backup to traditional liability, also serves as another potentially instructive case, demonstrating how industry-funded compensation systems can ensure swift remedy for associated harms.<sup>51</sup> In the case of AI harms generally and AI agent harms in particular, a no-fault compensation approach could serve a valuable function, but this would certainly depend, among other things, on establishing compensation schedules for different categories of harms—for example, financial losses from unauthorized transactions, business disruption from erroneous AI agent activity, etc.—providing clarity for both victims and fund administrators.<sup>52</sup>

The core advantage of a no-fault compensation approach to addressing AI agent harms is that it circumvents factors which might otherwise impede conventional fault-based approaches such as tort liability. AI supply chain complexities, opaque AI agent design choices, and the potential for harm, even with state-of-the-art safeguards, may render conventional liability pathways both slow and ineffective. Rather than subject victims to navigating this terrain, no-fault compensation schemes need only verify that harm occurred and that it falls within covered categories. A further advantage of a no-fault compensation approach is its provision of predictable compensation timelines that benefit both victims and industry participants. By design, no-fault systems prioritize administrative efficiency over fault adjudication, channeling claims through streamlined processes rather than adversarial litigation. For AI agent harms, this structural difference could significantly reduce compensation timelines while simultaneously providing developers with clearer cost visibility for planning and pricing their technologies.

These advantages notwithstanding, achieving effective no-fault compensation for AI agent harms must grapple with a number of practical impediments. For instance, the absence of sufficient relevant actuarial data constitutes a critical bottleneck.<sup>53</sup> Without prevalent and robust risk assessment data, fund administrators cannot accurately estimate appropriate contribution rates or compensation schedules, potentially leading to either underfunding or excessive industry costs that risk stifling innovation. Furthermore, unlike automobile

---

<sup>50</sup> Andrew Hurst, “What Does No-Fault State Mean?” Policygenius, June 13, 2023, <https://www.policygenius.com/auto-insurance/states-with-no-fault-insurance/>; Kimberly M. Thompson, “Performance of the United States Vaccine Injury Compensation Program (VICP): 1988–2019,” *Vaccine* 38, no 9 (February 24, 2020): 2136-2143, <https://doi.org/10.1016/j.vaccine.2020.01.042>.

<sup>51</sup> United States Coast Guard, National Pollution Funds Center, “Oil Spill Liability Trust Fund (OSLTF),” accessed August 8, 2025, [https://www.uscg.mil/Mariners/National-Pollution-Funds-Center/About\\_NPFC/OSLTF/](https://www.uscg.mil/Mariners/National-Pollution-Funds-Center/About_NPFC/OSLTF/).

<sup>52</sup> For a discussion about insurance as a tool for general AI regulation, see Anat Lior, “Insuring AI: The Role of Insurance in Artificial Intelligence Regulation,” *Harvard Journal of Law & Technology* 35, no. 2 (September 12, 2023): 467-530, [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4266259](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4266259). For a discussion about the significance of liability insurance in particular (in contrast to no-fault insurance), see Kenneth S. Abraham and Catherine M. Sharkey, “The Glaring Gap in Tort Theory,” *Yale Law Journal* 133, no. 7 (Forthcoming): 2165-2255, <https://ssrn.com/abstract=4585790>.

<sup>53</sup> Martin Eling, “How insurance can mitigate AI risks,” Brookings, November 7, 2019, <https://www.brookings.edu/articles/how-insurance-can-mitigate-ai-risks/>.

accidents, vaccine-related injuries, or oil spills, which all occur within relatively predictable or well-understood parameters, AI agent harms can manifest across a range of deployment contexts.<sup>54</sup> Relevant actuarial data will need to suitably represent such diverse harm categories in order to facilitate a workable no-fault compensation scheme. Another potential challenge is the possibility of adverse selection when participation in a no-fault compensation scheme is voluntary. Under this model, developers with the safest systems may opt out to avoid subsidizing higher-risk competitors, leaving no-fault compensation funds dominated by companies with poorer safety records. This could lead to either unsustainably high contribution rates or inadequate compensation levels for victims. Finally, as for other kinds of insurance, a no-fault compensation scheme for AI agent harms could generate moral hazard by reducing developers' incentives to continuously ensure AI agents remain safe and reliable.<sup>55</sup> If companies know that harms will be compensated through collective funds rather than remediated through liability frameworks, they may systematically under-invest in safety research or reliability evaluations.

To help resolve the range of barriers impeding effective no-fault compensation for AI agent harms, policymakers can apply a mix of general policy solutions and targeted interventions. For example, the federal government could help drive the collection of data on AI agent failures and harms by instituting an incident reporting program, a proposal which has already received attention in the broader AI policy discourse.<sup>56</sup> This would help provide an actuarial foundation necessary for effective no-fault compensation schemes. No-fault compensation pilot programs for specific AI agent applications or sectoral deployments could also help iterate toward a more general scheme over the long term. Initial experimental implementations might focus on narrow domains such as AI agent use cases in financial services, automated administrative systems, or customer service-agents, where both risks and stakeholders are more clearly defined. For instance, a pilot program for AI financial agents could cover unauthorized transactions above minimum thresholds, providing clear boundaries while generating operational experience for broader schemes. Lastly, hybrid compensation approaches may prove more politically and practically feasible than blanket no-fault compensation. For example, no-fault automobile insurance systems preserve tort liability for severe injuries that exceed specific thresholds, such as monetary amounts or certain classes of injury.<sup>57</sup> The Oil Spill Liability Trust Fund illustrates another model of hybrid compensation, providing immediate compensation to victims when responsible parties cannot or will not pay, while preserving liability and actively seeking cost recovery from those parties afterward—thus maintaining deterrent effects through a temporal rather than threshold-based approach. AI agent compensation schemes could conceivably draw inspiration from these hybrid approaches, providing swift remedy for harms while preserving liability

---

<sup>54</sup> Renee Henson, "Government-Backed Insurance for Artificial Intelligence Technologies," *Georgia State University Law Review* 41, no. 3 (Spring 2025): 559-629, <https://dx.doi.org/10.2139/ssrn.5226107>.

<sup>55</sup> Tom Baker, "On the Genealogy of Moral Hazard," *Texas Law Review* 75, no. 2 (December, 1996): 237-296, [https://scholarship.law.upenn.edu/faculty\\_scholarship/872/](https://scholarship.law.upenn.edu/faculty_scholarship/872/); Ronen Avraham and Ariel Porat, "The Dark Side of Insurance," *Review of Law & Economics* 19, no. 1 (February, 2023): 13-45, <https://dx.doi.org/10.2139/ssrn.4203765>.

<sup>56</sup> John Croxton et al., "Establishing an AI Incident Reporting System," Federation of American Scientists, June 25, 2024, <https://fas.org/publication/establishing-an-ai-incident-reporting-system/>.

<sup>57</sup> M. Lee and Heather Newton, "No Fault Car Insurance: Overview," EBSCO Research Starters, 2018, <https://www.ebsco.com/research-starters/law/no-fault-car-insurance-overview>.

mechanisms through temporal, threshold-based, or other kinds of conditions.<sup>58</sup> This would help mitigate the risk of moral hazard by maintaining some measure of accountability and deterrence effects.

## FUTURE DIRECTIONS: HYBRID APPROACHES

The stick, carrot, and net approaches present policymakers with a range of pathways for addressing AI agent harms. Rather than representing mutually exclusive alternatives, these approaches could be strategically combined to leverage their distinct advantages and compensate for the particular challenges that each faces individually. In other contexts, there are analogues incorporating different aspects of the stick, carrot, or net approaches that support the potential of hybrid approaches. For instance, the cases of no-fault automobile insurance and the Oil Spill Liability Trust Fund illustrate how no-fault compensation schemes can operate alongside liability frameworks. Similarly, liability safe harbor proposals need not erect complete barriers to remediation. Limited liability shields—which may be configured in a number of different ways—could preserve some form of remediation and deterrence while still remaining enticing to developers and providing them with clear exposure limits. The carrot and net approaches could also conceivably be blended together, for instance, by conditioning liability immunity on both third party auditing and participation in no-fault compensation schemes that provide remedy for verified harms.

Ultimately, the challenge lies in designing policy interventions that harness the strengths of each approach while mitigating their respective limitations. The transparency measures that would help facilitate effective liability frameworks would also potentially support more accurate actuarial assessment for no-fault compensation schemes. The standardization that emerges from industry auditing programs could inform both judicial expectations of reasonable care and compensation schedules for no-fault AI insurance programs. Policymakers should consider how to thoughtfully calibrate interventions across the spectrum of stick, carrot, and net approaches to achieve outcomes at the optimal point between upholding the public interest and driving responsible AI innovation.

---

<sup>58</sup> For further discussion about the role of liability and no-fault insurance mechanisms for addressing a more general set of AI harms, see Gabriel Weil et al., “Insuring Emerging Risks from AI,” Oxford Martin School, November 19, 2024, <https://www.oxfordmartin.ox.ac.uk/publications/insuring-emerging-risks-from-ai>.

---

## REFERENCES

Abraham, Kenneth S., and Catherine M. Sharkey. "The Glaring Gap in Tort Theory." *Yale Law Journal* 133, no. 7 (May, 2024): 2165-2255.

<https://ssrn.com/abstract=4585790>

"Authorities Warn of TikTok Viral Video Challenge That's Causing Fires." *CBS News*. January 21, 2020.

<https://www.cbsnews.com/boston/news/penny-socket-charger-challenge-tiktok-fire-video/>

Avraham, Ronen, and Ariel Porat. "The Dark Side of Insurance." *Review of Law & Economics* 19, no. 1 (February, 2023): 13-45.

<https://dx.doi.org/10.2139/ssrn.4203765>

Baker, Tom. "On the Genealogy of Moral Hazard." *Texas Law Review* 75, no. 2 (December, 1996): 237-296.

[https://scholarship.law.upenn.edu/faculty\\_scholarship/872/](https://scholarship.law.upenn.edu/faculty_scholarship/872/)

Ball, Dean W. "How AI Liability Should Work (Part II)." *Hyperdimensional*. February 26, 2025.

<https://www.hyperdimensional.co/p/how-should-ai-liability-work-part-3df>

Ball, Dean W. "Putting Private AI Governance into Action." *Hyperdimensional*. March 20, 2025.

<https://www.hyperdimensional.co/p/putting-private-governance-into-action>

Bambauer, Derek E., and Melanie J. Teplinsky. "Shields Up For Software." *Lawfare*. December 19, 2023.

<https://www.lawfaremedia.org/article/shields-up-for-software>

Belcic, Ivan and Cole Stryker. "AI Agents in 2025: Expectations vs. Reality." *IBM Think*. March 4, 2025.

<https://www.ibm.com/think/insights/ai-agents-2025-expectations-vs-reality>

Brown, Ian. "Allocating Accountability in AI Supply Chains." *Ada Lovelace Institute*. June 29, 2023.

<https://www.adalovelaceinstitute.org/resource/ai-supply-chains/>

Buiten, Miriam, Alexandre de Streel, and Martin Peitz. "EU Liability Rules for the Age of Artificial Intelligence." *Centre on Regulation in Europe*. March 18, 2021.

<https://cerre.eu/publications/eu-liability-rules-age-of-artificial-intelligence-ai/>

Buiten, Miriam, Alexandre de Streel, and Martin Peitz. "The Law and Economics of AI Liability." *Computer Law & Security Review* 48, no. 1 (April, 2023): 105794.

<https://doi.org/10.1016/j.clsr.2023.105794>

California Legislature. Senate Bill No. 813. Multistakeholder Regulatory Organizations. 2025–2026 Regular Session (CA 2025). Amended May 1, 2025. Accessed May 14, 2025.

---

<https://legiscan.com/CA/text/SB813/2025>

Calvert, Clay. “Suing OpenAI for ChatGPT-Produced Defamation: A Futile Endeavor?” American Enterprise Institute. January 8, 2025.

<https://www.aei.org/technology-and-innovation/suing-openai-for-chatgpt-produced-defamation-a-futile-endeavor/>

Casper, Stephen, David Krueger, and Dylan Hadfield-Menell. “Pitfalls of Evidence-Based AI Policy.” *arXiv:2502.09618*. April 18, 2025.

<https://doi.org/10.48550/arXiv.2502.09618>.

Chen, Pin-Yu. “What is AI adversarial robustness?” IBM Research. December 11, 2021.

<https://research.ibm.com/blog/securing-ai-workflows-with-adversarial-robustness>

Choi, Bryan H. “Negligence Liability for AI Developers.” Lawfare. September 26, 2024.

<https://www.lawfaremedia.org/article/negligence-liability-for-ai-developers>

“Contributory Negligence/Comparative Fault Laws in all 50 States.” Matthiesen, Wickert & Lehrer, S.C.. June 4, 2025.

<https://www.mwl-law.com/wp-content/uploads/2018/02/4413916.pdf>.

Croxton, John, David Robusto, Satya Thallam, and Doug Calidas. “Establishing an AI Incident Reporting System.” Federation of American Scientists. June 25, 2024.

<https://fas.org/publication/establishing-an-ai-incident-reporting-system/>

Diamantis, Mihailis. “Reasonable AI: A Negligence Standard.” *Vanderbilt Law Review* 77, no. 2 (forthcoming): 1-39.

<https://dx.doi.org/10.2139/ssrn.4609545>

Edwards, Benj. “Company apologizes after AI support agent invents policy that causes user uproar.” *Wired*. April 19, 2025.

<https://www.wired.com/story/cursor-ai-hallucination-policy-customer-service/>

Eling, Martin. “How insurance can mitigate AI risks.” Brookings. November 7, 2019.

<https://www.brookings.edu/articles/how-insurance-can-mitigate-ai-risks/>

Farid, Hanie, and Brandie Nonnecke. “The Case for Regulating Platform Design.” *Wired*, March 13, 2023.

<https://www.wired.com/story/make-platforms-safer-regulate-design-section-230-gonzalez-google/>

Fried, Ina. “OpenAI and Visa prep for AI-powered shopping.” *Axios*. May 1, 2025.

<https://www.axios.com/2025/05/01/chatbots-shopping-visa-ai-payments>

Fowler, Geoffrey A. "I let ChatGPT's new 'agent' manage my life. It spent \$31 on a dozen eggs." *The Washington Post*, February 7, 2025.

<https://www.washingtonpost.com/technology/2025/02/07/openai-operator-ai-agent-chatgpt/>

Gabriel, Iason, Geoff Keeling, Arianna Manzini, and James Evans. "We need a new ethics for a world of AI agents." *Nature* 644, no. 1 (August 4, 2025): 38-40.

<https://doi.org/10.1038/d41586-025-02454-5>

Goldman, Sharon. "Exclusive: Ex-Meta AI leaders debut an agent that scours the web for you in a push to ultimately give users their own digital 'chief of staff'." *Fortune*. June 10, 2025.

<https://fortune.com/2025/06/10/exclusive-ex-meta-ai-leaders-agent-web-yutori/>

Gray, Conrad. "Grok 4—The Good, The Bad and The Ugly - Sync #527." *Humanity Redefined*. July 13, 2025.

<https://www.humanityredefined.com/p/sync-527>

Haykel, Iskandar. "AI Security Tax Incentives." Americans for Responsible Innovation. March 27, 2025.

<https://ari.us/policy-bytes/ai-security-tax-incentives/>

Heikkilä, Melissa. "What Are AI Agents? Here's Everything You Need to Know." *MIT Technology Review*. July 5, 2024.

<https://www.technologyreview.com/2024/07/05/1094711/what-are-ai-agents/>

Henson, Renee. "Government-Backed Insurance for Artificial Intelligence Technologies." *Georgia State University Law Review* 41, no. 3 (Spring 2025): 559-629.

<https://dx.doi.org/10.2139/ssrn.5226107>

Henson, Renee. "I am Become Death, the Destroyer of Worlds': Applying Strict Liability to Artificial Intelligence as an Abnormally Dangerous Activity." *Temple Law Review* 93, no. 3 (July 15, 2024): 349-390.

<https://dx.doi.org/10.2139/ssrn.4894986>

Hurst, Andrew. "What Does No-Fault State Mean?" Policygenius. June 13, 2023.

<https://www.policygenius.com/auto-insurance/states-with-no-fault-insurance/>

Krapayoon, Jam, Zoe Williams, and Rida Fayyaz. "AI Agent Governance: A Field Guide." Institute for AI Policy and Strategy. April 17, 2025.

<https://www.iaps.ai/research/ai-agent-governance>

Kwa, Thomas, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. "Measuring AI Ability to Complete Long Tasks." *arXiv:2503.14499*. March 30, 2025.

---

<https://doi.org/10.48550/arXiv.2503.14499>.

Kwon, Joe. “Democracy in the Dark: Why AI Transparency Matters.” TechPolicy Press. May 9, 2025.  
<https://www.techpolicy.press/democracy-in-the-dark-why-ai-transparency-matters/>

“Leave it to Manus: Benchmarks,” Manus AI, accessed August 8, 2025.  
<https://manus.im/home>

Lee, M., and Heather Newton. “No Fault Car Insurance: Overview.” EBSCO Research Starters. 2018.  
<https://www.ebsco.com/research-starters/law/no-fault-car-insurance-overview>

Levine, Dan, and Hyunjoon Jin. “Tesla wins Autopilot trial involving fatal crash.” *Reuters*. November 1, 2023.  
<https://www.reuters.com/business/autos-transportation/tesla-wins-autopilot-trial-involving-fatal-crash-2023-10-31/>

Lior, Anat. “Holding AI Accountable: Addressing AI-Related Harms Through Existing Tort Doctrines.” Symposium: How AI Will Change the Law, *University of Chicago Law Review* (2024).  
<https://lawreview.uchicago.edu/online-archive/holding-ai-accountable-addressing-ai-related-harms-through-existing-tort-doctrines>

Lior, Anat. “Insuring AI: The Role of Insurance in Artificial Intelligence Regulation.” *Harvard Journal of Law & Technology* 35, no. 2 (November, 2022): 467-530.  
[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4266259](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4266259)

Marcus, Gary. “AI Agents have, so far, mostly been a dud.” *Marcus on AI*. August 3, 2025.  
<https://garymarcus.substack.com/p/ai-agents-have-so-far-mostly-been>

Marshall, Aarian. “Tesla Found Partly Liable in 2019 Autopilot Death.” *Wired*. August 1, 2025.  
<https://www.wired.com/story/tesla-liable-2019-autopilot-crash-death/>

Marshall, Aarian. “Why Wasn’t Uber Charged in a Fatal Self-Driving Car Crash?” *Wired*. September 17, 2020.  
<https://www.wired.com/story/why-not-uber-charged-fatal-self-driving-car-crash/>

Masse, Bryson. “A Chevy for \$1? Car dealer chatbots show AI perils.” *VentureBeat*. December 19, 2023.  
<https://venturebeat.com/ai/a-chevy-for-1-car-dealer-chatbots-show-perils-of-ai-for-customer-service/>

Mialon, Grégoire, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. “GAIA: a benchmark for General AI Assistants.” *arXiv:2311.12983*. November 21, 2023.  
<https://doi.org/10.48550/arXiv.2311.12983>

Millière, Raphaël. “Normative Conflicts and Shallow AI Alignment.” *arXiv:2506.04679*. June 5, 2025.  
<https://doi.org/10.48550/arXiv.2506.04679>

Mitchell, Margaret, Avijit Ghosh, Sasha Luccioni, and Giada Pistilli. “Why Handing Over Total Control to AI Agents Would Be a Huge Mistake.” *MIT Technology Review*. March 24, 2025.

<https://www.technologyreview.com/2025/03/24/1113647/why-handing-over-total-control-to-ai-agents-would-be-a-huge-mistake/>

Muppidi, Sri. “OpenAI Forecasts Revenue Topping \$125 Billion in 2029 as Agents, New Products Gain.” *The Information*. April 23, 2025.

<https://www.theinformation.com/articles/openai-forecasts-revenue-topping-125-billion-2029-agents-new-products-gain>

Nassauer, Sarah, and Jacob Gershman. “Activist Robby Starbuck Sues Meta Over AI Answers About Him.” *Wall Street Journal*. April 29, 2025.

<https://www.wsj.com/tech/ai/activist-roby-starbuck-sues-meta-over-ai-answers-about-him-9eba5d8a>

Porter, Zoe, Annette Zimmermann, Phillip Morgan, John McDermid, Tom Lawton, and Ibrahim Habli. “Distinguishing Two Features of Accountability for AI Technologies.” *Nature Machine Intelligence* 4, no. 1 (September 22, 2022): 734-736.

<https://doi.org/10.1038/s42256-022-00533-0>

Price II, W. Nicholson, and I. Glenn Cohen. “Locating Liability for Medical AI.” *DePaul Law Review* 73, no. 2 (Spring 2024): 339-368.

<https://dx.doi.org/10.2139/ssrn.4517740>

Rein, David, Joel Becker, Amy Deng, Seraphina Nix, Chris Canal, Daniel O’Connel, Pip Arnott, Ryan Bloom, Thomas Broadley, Katharyn Garcia, Brian Goodrich, Max Hasin, Sami Jawhar, Megan Kinniment, Thomas Kwa, Aron Lajko, Nate Rush, Lucas Jun Koba Sato, Sydney Von Arx, Ben West, Lawrence Chan, and Elizabeth Barnes. “HCAST: Human-Calibrated Autonomy Software Tasks.” *arXiv:2503.17354*. March 21, 2025.

<https://doi.org/10.48550/arXiv.2503.17354>

Responsible Innovation and Safe Expertise Act of 2025 (RISE Act of 2025). S. 2081. 119th Cong., 1st sess. Introduced by Sen. Cynthia M. Lummis. 2025.

<https://www.lummis.senate.gov/wp-content/uploads/Rise-Act-Text.pdf>

Robusto, David. “The Technologist-Economist Disconnect: Assessing AI Labor Disruption.” Americans for Responsible Innovation. April 7, 2025.

<https://ari.us/wp-content/uploads/2025/04/Report-Technologist-Economist-Disconnect-ARI04072025-1.pdf>

Rosenzweig, Paul. “Content Moderation and the Least Cost Avoider.” *Lawfare*. June 10, 2024.

<https://www.lawfaremedia.org/article/content-moderation-and-the-least-cost-avoider>

Sabin, Sam. “Exclusive: Anthropic warns fully AI employees are a year away.” *Axios*. April 22, 2025.

---

<https://www.axios.com/2025/04/22/ai-anthropic-virtual-employees-security>

Sapkota, Ranjan, Konstantinos I. Roumeliotis, and Manoj Karkee. "AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges." arXiv preprint *arXiv:2505.10468v4*. May 28, 2025.

<https://doi.org/10.48550/arXiv.2505.10468>

Selbst, Andrew D. "Negligence and AI's Human Users." *100 Boston University Law Review* 1315 (January 28, 2020): 1315-1376.

<https://ssrn.com/abstract=3350508>

Sharkey, Catherine. "Products Liability for Artificial Intelligence." Lawfare. September 25, 2024.

<https://www.lawfaremedia.org/article/products-liability-for-artificial-intelligence>

Shavit, Yonadav, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Alex Beutel, Alexandre Passos, and David G. Robinson. "Practices for Governing Agentic AI Systems." OpenAI. December 14, 2023.

<https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf>

Shead, Sam. "Amazon's Alexa Assistant Told a Child to Do a Potentially Lethal Challenge." *CNBC*, December 29, 2021.

<https://www.cnn.com/2021/12/29/amazons-alexa-told-a-child-to-do-a-potentially-lethal-challenge.html>

Shojaee, Parshin, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity." Apple Machine Learning Research. May 28, 2025.

<https://machinelearning.apple.com/research/illusion-of-thinking>

Siegel, Zachary S., Sayash Kapoor, Nitya Nagdir, Benedikt Stroebel, and Arvind Narayanan. "CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark." *arXiv:2409.11363*. September 17, 2024.

<https://doi.org/10.48550/arXiv.2409.11363>

Skarlinski, Michael, Tyler Nadolski, James Braza, Remo Storni, Mayk Caldas, Ludovico Mitchener, Michaela Hinks, Andrew White, and Sam Rodrigues. "Superintelligent AI Agents for Scientific Discovery." FutureHouse. May 1, 2025.

<https://www.futurehouse.org/research-announcements/launching-futurehouse-platform-ai-agents>

Smith, Gregory, Karlyn D. Stanley, Krystyna Marcinek, Paul Cormarie, and Salil Gunashekar. "Liability for Harms from AI Systems: The Application of U.S. Tort Law and Liability to Harms from Artificial Intelligence Systems." RAND Corporation. November 20, 2024.

[https://www.rand.org/pubs/research\\_reports/RRA3243-4.html](https://www.rand.org/pubs/research_reports/RRA3243-4.html)

Thompson, Kimberly M. “Performance of the United States Vaccine Injury Compensation Program (VICP): 1988–2019.” *Vaccine* 38, no 9 (February 24, 2020): 2136-2143.

<https://doi.org/10.1016/j.vaccine.2020.01.042>

Tiku, Nitashi. “An AI companion suggested he kill his parents. Now his mom is suing.” *Washington Post*. December 13, 2024.

<https://www.washingtonpost.com/technology/2024/12/10/character-ai-lawsuit-teen-kill-parents-texas/>

Tocchetti, Andrea, Lorenzo Corti, Agathe Balayn, Mireia Yurrita, Philip Lippmann, Marco Brambilla, and Jie YangTocchetti. “A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities.” *ACM Computing Surveys* 57, no. 6 (February 10, 2025): 1-38.

<https://doi.org/10.1145/3665926>

Tomei, Philip Moreira, Rupal Jain, and Matija Franklin. “AI Governance through Markets.” *arXiv:2501.17755*. March 5, 2025.

<https://doi.org/10.48550/arXiv.2501.17755>

United States Coast Guard, National Pollution Funds Center. “Oil Spill Liability Trust Fund (OSLTF).” accessed August 8, 2025.

[https://www.uscg.mil/Mariners/National-Pollution-Funds-Center/About\\_NPFC/OSLTF/](https://www.uscg.mil/Mariners/National-Pollution-Funds-Center/About_NPFC/OSLTF/)

Weil, Gabriel, Matteo Pistillo, Suzanne Van Arsdale, Junichi Ikegami, Kensuke Onuma, Megumi Okawa, and Michael A. Osborne. “Insuring Emerging Risks from AI.” Oxford Martin School. November 19, 2024.

<https://www.oxfordmartin.ox.ac.uk/publications/insuring-emerging-risks-from-ai>

Weil, Gabriel. “The Case for AI Liability.” *AI Frontiers*. June 12, 2025.

<https://ai-frontiers.org/articles/case-for-ai-liability>

Weil, Gabriel. “Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence.” *SSRN Scholarly Paper* no. 4694006 (June 6, 2024): 1–80.

<https://dx.doi.org/10.2139/ssrn.4694006>

Wolfe, Cameron R. “AI Agents from First Principles.” *Deep (Learning) Focus*. June 9, 2025.

<https://cameronwolfe.substack.com/p/ai-agents>

Yagoda, Maria. “Airline held liable for its chatbot giving passenger bad advice - what this means for travellers.” *BBC*. February 23, 2024.

<https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know>

Yang, Angela. “Lawsuit claims Character.AI is responsible for teen’s suicide.” *NBC News*. October 23, 2024.

---

<https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791>