Article

Cecil Abungu*

Foreseeing the Unforeseeable: How U.S. Negligence Law Should Address the Foreseeability of Harms Caused by Autonomous AI Agents

https://doi.org/10.1515/jtl-2025-0027
Received August 13, 2025; accepted March 9, 2026; published online April 7, 2026

Abstract: As AI systems increasingly perform tasks with limited human oversight, courts will soon be required to determine how negligence law should respond when autonomous AI agents cause personal injury. Traditional foreseeability doctrine, as many scholars have observed, may fail to account for the opacity and unpredictability that characterize these systems. The main challenge could arrive when AI developers claim that a harmful outcome was unforeseeable because the specific causal pathway was novel, complex, or obscure. This article argues that such reasoning misallocates responsibility. Building on recent scholarly work, it takes the view that opacity and unpredictability are not inherent features of advanced AI systems. Rather, they are the result of abstraction choices made by AI developers, who often prioritize accuracy and efficiency over interpretability and predictability. When those choices increase the likelihood of opaque or unexpected outcomes, the legal framework should be adjusted to reflect that responsibility. In particular, where the foreseeability standard in negligence law typically makes it difficult for plaintiffs to prevail, it should be relaxed in their favor. The article examines how U.S. courts apply foreseeability across duty, breach, and proximate cause, and identifies the duty stage as the most urgent point for reform. It proposes a three-part doctrinal framework for cases involving personal injury caused by autonomous AI agents. First, courts should preserve existing law where foreseeability is already sympathetic to plaintiffs. Second, they should replace overly fact-intensive duty inquiries with clear, plaintiff-friendly categorical reasoning. Third, they should retain fact-intensive analysis at the breach and proximate cause stages to prevent

I am grateful to Elvis Mogesa Ongiri for his exceptional research assistance.

***Corresponding author: Cecil Abungu**, PhD candidate, University of Cambridge, Cambridge, UK; Coordinator, ILINA Program; and Research Affiliate, Institute for Law and AI, E-mail: cecil.abungu@law-ai.org

overextension of liability. This approach maintains foreseeability as a meaningful constraint while calibrating it to the distinctive risks posed by autonomous AI agents.

Keywords: autonomous AI agents; negligence; foreseeability; abstraction choices

1 Introduction

Imagine a hospital that relies on an autonomous AI agent to schedule patient treatments and assign operating rooms. The agent erroneously deprioritizes a critical cardiac patient for surgery, resulting in the patient's death. In another scenario, a cloud-based autonomous agent is used to control home devices, including thermostats, kitchen appliances, and an electric stove. While the resident is asleep, the agent activates the stove, igniting a nearby flammable item and causing serious injuries from burns and smoke inhalation.

These hypotheticals illustrate the possibility that autonomous AI agents could soon be capable of causing personal injury. Once confined to research labs and science fiction, such agents are now within sight. In 2025, researchers allowed an AI system to run an automated retail store for an entire month, signaling that agents operating independently in the physical world are not merely speculative.[1] More recently, OpenAI introduced an autonomous agent designed to complete a broad range of computer-based tasks without direct supervision, further demonstrating that these systems are poised to operate in increasingly complex and consequential settings.[2] The accelerating development of autonomous agents holds significant promise, but it also raises urgent legal questions. Chief among them is this: when an autonomous AI agent causes physical harm, who should be held responsible under the law?

Autonomous AI agents are increasingly capable of making complex decisions without real-time human oversight. These systems can perceive their environment, evaluate options, act, and learn from past outcomes.[3] Their autonomy offers substantial benefits, including greater efficiency, consistent performance, and the ability to complete tasks that would be too time-consuming, dangerous, or complex for

1 Anthropic & Andon Labs, *Project Vend: Can Claude Run a Small Shop? (And Why Does That Matter?)*, (June 27, 2025), https://www.anthropic.com/research/project-vend-1 [https://perma.cc/BH26-SAD9] (last visited July 31, 2025).

2 OpenAI, *Introducing ChatGPT Agent: Bridging Research and Action* (July 17, 2025), https://openai.com/index/introducing-chatgpt-agent/ (last visited July 19, 2025).

3 Margaret Michelle et al., *Fully Autonomous AI Agents Should Not Be Developed*, arXiv, 1, 1 (Feb. 4, 2025), arXiv:2502.02649 [cs.AI], https://arxiv.org/abs/2502.02649.

humans to manage on their own.[4] At the same time, this independence introduces significant risks. By design, autonomous agents operate with a degree of discretion that makes their behavior difficult to predict. Unlike traditional rule-based software, modern AI systems, particularly those built using deep learning, can generate results that even their creators do not fully understand or anticipate.[5]

The internal workings of these systems often function as black boxes, with decision pathways that are opaque to human observers. As autonomous agents adapt to new data and interact with dynamic environments, they may behave in ways that are unexpected or unintuitive.[6] Developers may not program specific errors, but the emergent behavior of these systems can nonetheless result in harmful outcomes. This unpredictability, which enables them to discover novel solutions and improve over time, also means they may cause harm in ways no human explicitly foresaw.[7] When those harms involve physical injury or death, they present distinct challenges for the legal system, which must determine how to assign responsibility for decisions made by machines operating beyond direct human control.

That challenge comes into sharp focus in the tort law doctrine of negligence, which remains a primary vehicle for injured persons to seek redress. A key cornerstone of negligence law is the concept of foreseeability, which requires that a defendant is only liable if the plaintiff's injury was a foreseeable result of the defendant's breach of duty.[8] In traditional cases, foreseeability serves as a limit on liability on the basis that it would be unfair to hold someone accountable for freak accidents no one could have predicted.[9] But when an autonomous AI agent is the actor directly causing harm, foreseeability becomes a vexing question. AI developers and operators who are sued will argue that they exercised reasonable care and that the harmful outcome was beyond what they could have imagined. Indeed, AI's inscrutable, unintuitive decision processes can make specific errors essentially impossible to anticipate.

In a negligence lawsuit over an agent-induced injury, this creates a serious hurdle. The key question will be, was the harm a reasonably foreseeable consequence of the developer's conduct, or an unforeseeable anomaly? Plaintiffs will frame the risk in general terms whereas defendants may focus on the particular

4 Yonadav Shavit et al., *Practices for Governing Agentic AI Systems*, OpenAI, 7 (Dec. 2023), https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

5 Alan Chan et al., *Harms from Increasingly Agentic Algorithmic Systems*, *in* FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency 651, 657–58 (2023).

6 Maarten Herbosch, *Liability of AI Agents*, 26 N. C. J.L. & Tech. 391, 402–03 (2025).

7 Jason Wei et al., *Emergent Abilities of Large Language Models*, arXiv, 1, 1–2 (Oct. 26, 2022), arXiv:2206.07682 [cs.CL], https://arxiv.org/pdf/2206.07682.

8 W. Jonathan Cardi, *Purging Foreseeability*, 58 Vand. L. Rev. 739, 743–44 (2005).

9 Racheal Mulheron, Principles of tort law 51 (2d ed. Cambridge Univ. Press 2020).

chain of events as too remote or novel. In practical terms, if a court deems the AI's harmful misbehavior unforeseeable, a negligence claim will fail. Thus, the very features that make AI powerful (its autonomy and learning ability) could become a shield against liability under current foreseeability tests. How then should negligence law's foreseeability analysis apply when an autonomous AI agent causes personal injury?

This article tackles that problem and does so within a particular scope. First, the focus is on level 3 and 4 autonomous AI agents. These are highly advanced AI systems capable of performing tasks with minimal (Level 3) or effectively no (Level 4) human oversight.[10] By concentrating on levels 3–4, the discussion centers on agents that are truly technically autonomous, as opposed to mere automated tools requiring constant supervision. Second, the type of harm under consideration is personal injury as opposed to mere economic loss. This is because injuries to life, bodily integrity, or mental well-being have long received special solicitude in the law, often outweighing competing commercial interests.[11] They also present the most visceral and compelling test cases for AI liability. Third, the jurisdictional focus is on United States negligence law. The U.S. is home to the leading AI developers and has a well-developed body of tort doctrine, making it a natural proving ground for these issues. While the analysis may hold lessons for other common-law jurisdictions, the discussion will use U.S. legal principles and terminology.

At this point, a note on the choice of doctrinal framework is warranted. One might wonder why this article focuses on negligence rather than product liability, given that courts and policymakers seem increasingly open to treating AI systems as products.[12] The choice to focus on negligence is based on three reasons. First, negligence is the most doctrinally developed and widely available cause of action for personal injury in U.S. law, and it applies regardless of whether an AI system is treated as a product. Second, the foreseeability doctrine that is this article's focus operates distinctively within the negligence framework and does not play the same structural role in product liability's design-defect or failure-to-warn analyses. Third,

10 Mitchell et al., *supra* note 3, at 3.

11 Dilan C. Esper & Gregory C. Keating, *Putting Duty in Its Place: A Reply to Professors Goldberg and Zipursky*, 41 Loy. L.A. L. Rev. 1225, 1239 (2008).

12 *Garcia v. Character Technologies, Inc*., No. 6:24-cv-1903-ACC-DCI, 2025 WL 3456789, 32–36 (M.D. Fla. May 21, 2025); Karni A. Chagal-Feferkorn, *Am I an Algorithm or a Product?: When Products Liability Should Apply to Algorithmic Decision-Makers*, 30 Stan. L. & Pol'y Rev. 61, 82–84 (2019); Maarten Herbosch, Liability of AI Agents, 26 N. C. J.L. & Tech. 391, 425–427 (2025); Catherine Sharkey, *A Products Liability Framework for AI*, 25 Colum. Sci. & Tech. L. Rev. 240, 224 (2024); Raqda Sayidali, *What the* Megan Garcia *Case Can Tell Us About AI Liability in the U.S*., RAILS, (Nov. 3, 2025), https://blog.ai-laws.org/what-the-megan-garcia-case-tells-us-about-ai-liability-in-the-u-s/?cn-reloaded=1 last visited Feb. 23 2026.

even if product liability becomes the dominant pathway for AI injury claims, negligence claims will remain available in parallel, and hence the doctrinal questions addressed in this article will retain their relevance.[13]

Within the scope explained above, the core research questions are as follows. Does the foreseeability requirement in traditional negligence framework need to be adjusted for injuries caused by autonomous AI agents? If so, why is such adjustment necessary, and how should it be accomplished?

This article argues that adjustment to the foreseeability framework is indeed required. It proposes specific doctrinal adjustments, with a focus on the duty-foreseeability inquiry, that would ease the burden on plaintiffs in agent-related personal injury cases where an AI developer is sued. The proposed adjustment does not eliminate foreseeability from negligence analysis, nor does it advocate for unlimited liability. Rather, it targets the threshold at which foreseeability operates and proposes modest reforms to ensure that plaintiffs are not barred from recovery based on overly particularistic applications of the doctrine. The main adjustment proposed is that in U.S. states that still rely heavily on foreseeability at the duty stage, the law should permit courts to recognize broader categories of foreseeable harm. For example, courts should consider it foreseeable that autonomous agents deployed by developers may cause physical injury through misalignment, malfunction, or misuse.

The normative basis for this proposal is that opacity and autonomy in AI are not inevitable features of technology but are instead the result of abstraction choices made by AI developers. Developers routinely design systems to maximize performance at the expense of interpretability, and in doing so deliberately relinquish direct control to allow agents to operate independently.[14] These design decisions shape the risk profile of AI systems, and they should inform the structure of the legal standards that govern accountability. It would be inequitable to allow developers to avoid liability by invoking unpredictability that results from their own engineering decisions. Accordingly, foreseeability doctrine should be recalibrated to account for these abstraction choices. The goal is to ensure that injured plaintiffs are not denied relief simply because an autonomous AI agent behaved in an unexpected way. By anchoring the analysis in foreseeable categories of harm rather than specific

13 It is worth noting that the abstraction-choices argument developed in this article could potentially be mapped onto AI product liability analyses as well. However, a full examination of how the argument would work in product liability doctrine related to autonomous AI agents is beyond the scope of this article.

14 Andrew D. Selbst, Suresh Venkatasubramanian & I. Elizabeth Kumar, *Deconstructing Design Decisions: Why Courts Must Interrogate Machine Learning and Other Technologies*, 85 Ohio St. L. J. 415, 424–429 (2024).

mechanisms, courts can maintain a principled approach to liability that promotes both justice for claimants and accountability for developers.

In advancing this argument, the article builds on a growing body of work exploring tort liability for AI. Some scholars have written about the role that tort law can play when hazards are emerging and incompletely understood,[15] others have written about situations in which tort law can serve as an effective instrument for governing AI systems,[16] and others have written about the possibility of building new kinds of tort regimes for autonomous weapons.[17] This article joins these strands of work in attempting to decipher how the tort of negligence can be adapted to respond to harms caused by autonomous agents.

Scholars are also recognizing that foreseeability is a pivotal issue in tort law-AI questions. For example, Gabriel Weil observes that while some AI failures are relatively predictable, other failure modes, such as complex alignment failures or the misuse of AI by malicious actors, introduce profound uncertainty into foreseeability analysis.[18] Courts might disagree on whether a particular injury scenario was foreseeable, given the novelty of AI behavior. Weil thus suggests that judges should adopt a capacious conception of foreseeability and hold AI developers liable if the general risk was reasonably anticipated even if the specific harm was not.[19] It is notable that Weil's suggestion is specifically targeted at the proximate cause stage of the negligence framework.[20]

Maarten Herbosch similarly argues that the advent of AI, with its complexity and opacity, does not render negligence law obsolete.[21] He maintains that as long as the category of harm was foreseeable to a reasonable AI developer or user, liability can still be established.[22] In his law-and-economics analysis, Herbosch concludes that AI's unpredictability calls for targeted refinements, not a wholesale overhaul, of tort doctrine.[23] In other words, traditional principles remain sound; they only

15 Catherine M. Sharkey, *Common Law Tort as a Transitional Regulatory Regime: A New Perspective on Climate Change Litigation* in Climate Liberalism Perspectives on Liberty, Property and Pollution 104 (Jonathan H. Adler ed. 2023). See also, Catherine M. Sharkey, *Tort Law in the Age of Regulations*, 76 U. Toronto L.J. 74 (2026).

16 Mark A. Geistfeld, *A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation*, 105 Cal. L. Rev. 1611, 1621–22 (2017).

17 Rebecca Crootof, *War Torts: Accountability for Autonomous Weapons*, 164 U. Pa. L. Rev. 1347, 1386–89(2016).

18 Gabriel Weil, *Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence*, SSRN 1, 33–34, (Jan. 28, 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4694006.

19 *Id.* at 49–50.

20 *Id.*

21 Herbosch *supra* note 6, at 391.

22 *Id.* at 433–34.

23 *Id.* at 391.

need to be sensibly adapted to account for the fact that AI may make mistakes in surprising ways. This article joins those scholars in rejecting the notion that AI's black box nature should immunize its creators from negligence liability.

At the same time, some researchers have highlighted the uncertainty in how courts will apply foreseeability standards to AI. Matthew van der Merwe et al., for example, note that it is not clear how courts will determine whether a victim's injuries are unforeseeable in AI cases.[24] Different judges could reach markedly different conclusions about the same AI accident, as one court might readily find the harm foreseeable, while another might characterize it as an outlier beyond the scope of liability. These scholarly perspectives underscore both the urgency of the foreseeability issue and the range of approaches to resolving it. By engaging with this literature, the present article positions itself in the camp that believes tort law is fundamentally up to the task of governing AI, but only if courts consciously adapt doctrines like foreseeability to the unique context of autonomous AI agents.

While some of the highlighted scholars have persuasively argued that foreseeability should remain capacious in AI cases, their analyses share a common gap. They establish that AI unpredictability should not defeat negligence, but they do not comprehensively explain why the legal system should respond by expanding rather than contracting the scope of foreseeable harm. This article addresses that gap in one crucial way. It grounds the normative case for relaxing foreseeability in the concept of abstraction choices, demonstrating that the opacity and unpredictability of autonomous agents are identifiable features of developer decision-making rather than inherent features of the technology. This reframing transforms the doctrinal question from whether courts can accommodate AI unpredictability to how they should react to the unpredictability they engineered. Furthermore, whereas existing proposals have primarily focused on proximate cause or on negligence doctrine at a general level, this article locates the primary doctrinal intervention at duty stage, which is arguably the point at which structural obstacles to plaintiff recovery are most acute.

The remainder of this article is organized as follows. Part I provides background on autonomous AI agents, explaining what sets "Level 3" and "Level 4" AI agents apart and surveying their capabilities, potential benefits, and the risks they pose. This Part illustrates how these AI systems are built and why their operation can lead to unintended consequences. Part II examines the role of foreseeability in U.S. negligence law. It reviews how courts traditionally decide which harms are considered foreseeable and discusses how emerging AI scenarios strain these traditional concepts.

---

24 Matthew van der Merwe et al., *Tort Law and Frontier AI Governance*, Lawfare, (May 24, 2024), https://www.lawfaremedia.org/article/tort-law-and-frontier-ai-governance [https://perma.cc/57U5-ZHHH](last visited July 31, 2025).

Part III then advances the case for relaxing foreseeability tests in AI injury cases. It introduces the notion of developers' abstraction choices (design decisions that yield opaque, autonomous AI) and argues that these choices justify shifting the risk of unforeseeable harms onto the AI's creators. In this Part, the article reframes the black box problem as a policy reason to err on the side of finding foreseeability and also explains why the personal injury context is especially critical for evolving the law. Part IV outlines what a more plaintiff-friendly foreseeability standard could look like in practice. It considers doctrinal adjustments and addresses potential objections. This Part argues that U.S. negligence law can reasonably ease the plaintiff's burden on foreseeability without unleashing unlimited liability, by calibrating the test to ensure truly outlandish harms remain beyond liability. Part V carries the conclusion.

2 Autonomous AI Agents

Because this article is about autonomous AI agents, it is useful to begin by canvassing what they are and how they are built as well as their distinguishing features, benefits and risks. The sections that follow cover that ground.

2.1 What Autonomous AI Agents Are

In decoding what makes AI agentic, Margaret Mitchell et al. focus on autonomy as opposed to agency itself since any labels of agency can run into philosophical disputes around intentionality, and going around that debate seems essential because autonomous agents do not yet have mental states to ascertain their intentionality.[25] In their analysis, Mitchell et al. create a "sliding scale" taxonomy that begins from level zero to level four.[26] This taxonomy ("Mitchelle taxonomy") builds on work from other scholars who have taken a similar approach.[27]

25 Mitchell et al., *supra* note 3, at 3.

26 *Id.*

27 Sayash Kapoor et al., *AI agents that matter*, arXiv, 1, 1 (July 1, 2024), arXiv:2407.01502, https://arxiv.org/abs/2407.01502; Andrew Ng, (@AndrewYNg), X (June 13, 2024, 4:20 PM), https://x.com/AndrewYNg/status/1801295202788983136 [https://perma.cc/4G7N-NWSG] (last visited July 31, 2025); Corbu Grey, *5 levels of AI Agents*, Medium (Oct. 11, 2024), https://cobusgreyling.medium.com/5-levels-of-ai-agents-updated-0ddf8931a1c6 (last visited July 27, 2025); Nathan Lambert, *The AI Agent Spectrum*, (Dec. 18, 2024), https://www.interconnects.ai/p/the-ai-agent-spectrum [https://perma.cc/P8BB-STPR] (last visited July 27, 2025); Aymeric Roucher et al., *Introducing smolagents, a simple library to build agents*, (Dec. 31 2024), https://huggingface.co/blog/smolagents [https://perma.cc/2PNT-FPB7](last visited July 27, 2025).

In the Mitchell taxonomy, level 0 agents function as simple processors with no influence on decision-making or program logic. They execute tasks strictly as instructed, without autonomy or contextual understanding. All control lies with the human, who directs every step of the process.[28] Level 1 agents are different because they act as routers capable of selecting between predefined paths or functions based on input conditions. While there is some logic branching, humans still govern the system's behavior and decision points.[29] Level 2 agents, on the other hand, are tool callers. In other words, they have the ability to decide which functions or tools they should use to achieve a goal. Although they demonstrate increased autonomy in tool selection, humans still control the broader task definition and sequencing at play.[30]

The Mitchell taxonomy delineates level 3 agents as multi-step agents, meaning that they can manage sequences of actions and coordinate multiple stages of task execution. While humans define the overall objective, the agents in question have the ability to determine how and when actions should occur.[31] Finally, level 4 agents are fully autonomous, implying that they can independently generate and execute their own code. They determine the what, how, and when without human prompting, effectively removing humans from the decision-making loop (Figure 1).[32]

It appears that other scholars are referring to level 3 or level 4 autonomous AI agents when they write about "AI agents". For example, Noam Kolt defines AI agents as models that can plan and do highly complex tasks in a digital environment without human supervision,[33] Maarten Herbosch defines them as systems capable of operating independently with minimal or no human oversight[34] and Cullen O'Keefe et al. define them as "AI systems that can perform computer-based tasks as competently as human experts."[35] To facilitate clear analysis, this article adopts the Mitchell taxonomy, and the research contained here is specifically about both level 3 and level 4 agents as defined above. The article will use the term "autonomous AI agent", "autonomous agent" or "agent" in reference to level 3 and level 4 agents.

---

28 Mitchell et al., *supra* note 3, at 3 (explanation is adapted from Table 1).

29 *Id.*

30 *Id.*

31 *Id.*

32 *Id.*

33 Noam Kolt, *Challenges in Governing AI Agents*, Lawfare, (March, 3 2025), https://www.lawfaremedia.org/article/challenges-in-governing-ai-agents (last visited on July, 4 2025) Noam Kolt, *Governing AI Agents*, (101 Notre Dame L. Rev.) (forthcoming 2025) (manuscript at 3, 12–13) SSRN (Feb. 11, 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4772956.

34 Herbosch *supra* note 6, at 397.

35 Cullen O'Keefe et al., *Law-Following AI: Designing AI Agents to Obey Human* Laws, 94 Fordham L. R. (forthcoming 2025) (manuscript at 13), SSRN (June 13, 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5242643.

Agentic Level	Stars	Description	Term	Who's in Control
Level 0	☆☆☆☆	Model has no impact on decision-making or program logic. Executes tasks exactly as instructed.	Simple processor	Human: Full control (defines what, how, when)
Level 1	★☆☆☆	Model selects between pre-defined logic paths based on inputs. No task definition autonomy.	Router	Human: Controls the logic structure and outcomes
Level 2	★★☆☆	Model selects which tools/functions to call for a given input or goal.	Tool caller	Human: Defines the overall task and sequencing
Level 3	★★★☆	Model plans and coordinates sequences of actions across multiple steps to accomplish a goal.	Multi-step agent	Shared: Human defines the goal; system plans
Level 4	★★★★	Model autonomously generates and executes its own code, defines tasks, and manages execution.	Fully autonomous agent	System: Full autonomy (decides what, how, when)

Figure 1: Levels of autonomy in AI agents (Adapted from Mitchell et al.). *Id*.

Some examples of existing level 3 agents include Meta-developed *Cicero*, allegedly the first AI to achieve human-level performance in the strategy game Diplomacy by integrating natural language processing with strategic planning;[36] Google DeepMind's *Gato*, a generalist agent trained on diverse tasks, capable of performing over 600 activities, including playing video games and controlling robotic arms;[37] *Manus*, an AI agent designed to execute complex tasks such as resume

36 Meta, *CICERO: AI that can collaborate and negotiate with you*' 2022, https://about.fb.com/news/2022/11/cicero-ai-that-can-collaborate-and-negotiate-with-you/#:~:text=We%E2%80%99ve%20built%20CICERO%2C%20the%20first%20AI%20to%20play,in%20gameplay%20using%20strategic%20reasoning%20and%20natural%20language.; Noam Brown et al., *Human-Level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning*, 378 *Science* 1067, 1067–1074 (2022).
37 Scott Reed et al., *A Generalist Agent*, arXiv, 1, 1 (May 12, 2022; rev. Nov. 11, 2022), arXiv:2205.06175 [cs.AI], https://arxiv.org/pdf/2205.06175.

sorting and website creation without continuous human input;[38] OpenAI's recently announced *ChatGPT Agent*, a software agent that can autonomously operate across user interfaces to complete tasks such as booking flights and performing multi-step workflows with minimal supervision;[39] Anthropic's similarly recent Claude Code "agent teams," which can spin up and coordinate multiple sub-agents to complete branched, multi-step tasks (e.g., large codebase reviews) under a user-defined objective;[40] and Microsoft Copilot Studio autonomous agents, which can execute multi-step workflows by operating websites and desktop applications through the user interface when APIs are unavailable.[41] Although we are yet to see any level 4 agents publicly available, it would not be unreasonable to claim that this could change in the near future.

2.2 The Black Box Problem, Emergent Behaviors and Technical AI Autonomy

Like other deep learning models, autonomous agents are bedeviled by the so-called "black box problem". The deep neural networks at the core of building and running such models and agents consist of multiple interconnected layers processing vast amounts of data. This intricate structure renders the decision-making processes of the agents almost impossible to interpret. As a result, even developers struggle to understand or predict the rationale behind model outputs. Such opacity introduces significant concerns regarding uncertainty and unpredictability in AI-driven decisions.[42]

---

38 Severin Sorensen, *Manus AI: The Dawn of Autonomous Agents and What It Means for Business* (March 13, 2025), https://www.aretecoach.io/post/manus-ai-the-dawn-of-autonomous-agents-and-what-it-means-for-business#:~:text=Manus%20AI%20is%20a%20recently%20launched%20artificial%20intelligence,Butterfly%20Effect%29%2C%20a%20Chinese%20startup%20based%20in%20Wuhan. (last visited July 10, 2025).

39 OpenAI, *supra* note 2.

40 Anthropic, *Introducing Claude Opus 4.6* (Feb. 5, 2026), https://www.anthropic.com/news/claude-opus-4-6 last visited Feb. 23, 2026.

41 Charles Lamanna, *Unlocking Autonomous Agent Capabilities with Microsoft Copilot Studio*, Microsoft (Oct. 21, 2024), https://www.microsoft.com/en-us/microsoft-copilot/blog/copilot-studio/unlocking-autonomous-agent-capabilities-with-microsoft-copilot-studio/ last visited Feb. 23, 2026.

42 Arun Das and Paul Rad, *Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey*, arXiv, 1, 1 (June 23, 2020) arXiv:2006.11371 [cs.CV], https://arxiv.org/pdf/2006.11371; Alan Chan et al., V*isibility into AI' in* FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency 958, 960 (2024); Tim Miller, *Explanation in Artificial Intelligence: Insights From the Social Sciences*, 267 *Artificial Intelligence* 1, 1 (2019).

Of course, there are some technical efforts being made to resolve the black box problem. One of the most prominent, mechanistic interpretability, is the research effort to reverse-engineer the computations and representations inside trained neural networks into algorithms that are human-understandable to arrive at a "granular" and "causal" understanding.[43] Researchers in this field have successfully reverse-engineered small-scale neural networks and simple tasks,[44] and there has been some progress in larger neural networks.[45] Nevertheless, the field of mechanistic interpretability still faces a considerable mountain to climb. For example, researchers have yet to decipher a durable method to go around the challenge of polysemanticity, the phenomenon where individual neurons in large models encode multiple unrelated concepts simultaneously.[46] This makes the behavior of deep learning models, including autonomous agents, extremely difficult to interpret.

Apart from the black box problem, the emergent behavior problem is the second notable challenge that deep learning models (including autonomous agents) carry. When deep learning models scale in size and complexity, they often exhibit capabilities which appear suddenly and unpredictably.[47] Such capabilities seem to appear suddenly – rather than through gradual improvement – once models reach certain parameter thresholds. These makes deep learning models like autonomous agents unpredictable.[48] Such unpredictability poses clear risks because deep learning models and autonomous agents can find unintended shortcuts to pursue their objectives.[49]

Finally, technical AI autonomy refers to an AI system's ability to operate independently of direct human control by making decisions, adapting to new situations, and executing actions without real-time supervision.[50] In the context of Level 3 and Level 4 autonomous agents, it refers to the ability of such systems to plan and carry out multi-step tasks without human intervention (level 3), and to modify their own code or strategies in pursuit of a goal (level 4).[51] For example, when given a general instruction such as "help write a research paper," a Level 3 agent could divide the

43 Leonard Bereska & Efstratios Gavves, *Mechanistic Interpretability for AI Safety: A Review*, OpenReview, 1 (2024), https://openreview.net/pdf?id=ePUVetPKu6.

44 *Id.* at 10.

45 *Id.*

46 Xingyi Yang et al., *Mixture of Experts Made Intrinsically Interpretable*, arXiv, 1, 1–2 (Mar., 5 2025) arXiv:2503.07639 [cs.LG], https://arxiv.org/pdf/2503.07639.

47 Wei et al., *supra* note 7.

48 *Id.* at 3.

49 Bereska & Gavves, *supra* note 44 at 21, 24.

50 Kenneth R Walsh, Sathiadev Mahesh & Cherie C. Trumbach, *Autonomy in AI systems*, 41 The J. of Tech. St. 38, 42 (2021).

51 Mitchell et al., *supra* note 3, at 3 (explanation is adapted from Table 1).

task into subtasks such as searching the literature, drafting an outline, and writing sections, and then complete them without further guidance. A Level 4 agent could go further by developing new methods or code to improve performance or adapt to unforeseen challenges.[52]

This level of autonomy is made possible by certain design features. Autonomous agents are equipped with memory modules, tool-use capabilities such as web browsing or code execution, and internal planning loops that allow them to evaluate and adjust their behavior.[53] At Level 3, the agent can manage its workflow based on intermediate results. At Level 4, the agent may rewrite or generate new functions within its own architecture.[54] Technical AI autonomy carries significant risks, including the possibility that an agent will make harmful decisions without human awareness, adapt in unintended ways, or pursue goals misaligned with user intent.[55]

2.3 The Potential Benefits of Autonomous AI Agents

Autonomous agents can act as an 'impact multiplier' of the current benefits of AI systems,[56] which are already numerous. Consider some frequently-discussed case studies. Research has shown that when used in the natural sciences, agents can scour and synthesize vast literature collections, manage complex datasets, and automate the design and execution of experiments.[57] By translating research goals into Python or chemical-simulation code, these agents will be able to generate and test hypotheses much faster than human teams alone, dramatically shortening discovery cycles in fields like materials science or drug development.[58]

Autonomous agents are also expected to transform engineering workflows by taking over routine but complex tasks across software development, industrial

52 *Id.*

53 Lei Wang et al., *A Survey on Large Language Model based Autonomous Agents*, arXiv, 1, 3–4 (Dec., 15 2024), arXiv:2308.11432v6 [cs.AI], https://arxiv.org/abs/2308.11432.

54 Mitchell et al., *supra* note 3, at 3 (explanation is adapted from Table 1).

55 Yaniv Benhamou & Justine Ferland, *Artificial Intelligence and Damages: Assessing Liability and Calculating the Damages*, ResearchGate, 6–7 (2020) (forthcoming *in* Leading Legal Disruption: Artificial Intelligence and a Toolkit for Lawyers and the Law (Pina D'Agostino, Carole Piovesan & Aviv Gaon eds, Thomson Reuters Canada 2020)), ResearchGate, https://www.researchgate.net/publication/339140477_ARTIFICIAL_INTELLIGENCE_DAMAGES_ASSESSING_LIABILITY_AND_CALCULATING_THE_DAMAGES.

56 Shavit et al., *supra* note 4.

57 Reiichiro Nakano et al., *Webgpt: Browser-assisted question-answering with human feedback*, arXiv, 1, 1–2 (Dec., 17 2021), arXiv:2112.09332 [cs.CL], https://arxiv.org/pdf/2112.09332 (discussing AI systems ability to browse the web); Wang et al., *supra* note 54 at, 23–24.

58 Wang et al., *supra* note 54 at, 23–24.

automation, and robotics.[59] In manufacturing, for example, it is expected that agents will be able to integrate with digital twins to orchestrate adaptive production lines, optimize throughput and reduce downtime.[60] Concomitantly, in the field of robotics, it is expected that robots enmeshed with agents will be able to plan multi-step navigation tasks, learn from feedback, and coordinate sub-agents to tackle long-horizon objectives. This could lead to the creation of autonomous systems that can personalize household chores or execute precision assembly.[61] Advanced multi-agent systems can also offer benefits like decentralized, more democratic AI, improved coordination, greater robustness and efficiency, and new ways to address alignment and safety challenges.[62]

There are many other domains in which agents could positively transform human life. This is essentially what makes building and deploying them so attractive. However, alongside these benefits come significant risks, as this article shows in the next section.

2.4 The Risks of Autonomous AI Agents

Since autonomous agents are likely to be used in a wide-ranging way, they could cause many kinds of harm, from financial losses to physical and psychological injury.[63] The paragraphs that follow review some harms which the use of autonomous agents could result in.

Misuse harms could be the result of malicious or reckless use of autonomous agents.[64] As Chan and others have argued, agents can automate entire pipelines for harmful activities like designing bioweapons or running influence campaigns, making sophisticated attacks accessible to untrained individuals without human oversight.[65]

Malicious actors could use autonomous agents to infiltrate systems, exfiltrate data, or launch automated large-scale attacks, turning powerful tools into vectors for

59 Wenlong Huang et al., *Inner Monologue: Embodied Reasoning through Planning with Language Models*, arXiv, 1, 1–2 (July 12, 2022), arXiv:2207.05608 [cs.RO], https://arxiv.org/pdf/2207.05608 (discussing language models interactions with robots).

60 Wang et al., *supra* note 54 at, 24–25.

61 Wang et al., *supra* note 54 at, 24–27.

62 Lewis Hammond et al., *Multi-Agent Risks from Advanced AI*, arXiv, 1, 8 (Feb. 19, 2025), arXiv:2502.14143 [cs.MA], https://arxiv.org/abs/2502.14143.

63 Yangjun Ruan et al., *Identifying the Risks of LM-Agents with an LM-emulated Sandbox*, arXiv, 1, 1 (May 7, 2024), arXiv:2309.15817 [cs.AI], https://arxiv.org/abs/2309.15817.

64 Dan Hendrycks, Introduction to AI Safety, Ethics, and Society 6–9 (Taylor & Francis 2025).

65 Chan et al., *supra* note 43 at 959.

harm.[66] Malicious actors could equally exploit deep learning models' tendency to generate false or misleading content by using agents to tailor it to individuals, spreading it across larger platforms, and manipulating beliefs. This could enable large-scale deception, exploitation, and the amplification of harmful content like non-consensual intimate media.[67]

Autonomous agents could also lead to misalignment harms. These are understood to be harms that result from AI models showing unexpected or unintended behaviors and capabilities that are in conflict with human intentions or values.[68] Agents could lead to very damaging misalignment harms because their actions, while individually safe, can combine in unexpected and harmful ways. They may intentionally override or misinterpret guardrails,[69] especially in critical applications like medical or weapon systems, where failure could be life-threatening. As agents gain broader access to real-world domains, perform more complex tasks, and operate with less human oversight, their goals can increasingly diverge from human intent.[70] This risk is heightened when agents use human-like interfaces to perform actions indistinguishable from real users without triggering detection.[71] Moreover, LLM-agents tend to display power-seeking behavior when operating within text-based adventure game environments.[72]

Finally, when AI Agents create specialized sub-agents to tackle subtasks, oversight becomes harder and the number of potential failure points multiplies, as each new sub-agent carries its own misalignment risk.[73] In recognition of this risk of implosion, the United Nations High-level Advisory Board on AI has called on the governance against potential "loss of human control over autonomous agents".[74]

---

66 Maximilian Mozes et al., *Use of LLMs for Illicit Purposes: Threats, Prevention Measures, and Vulnerabilities*, arXiv, 1, 6–8 (Aug. 24, 2023), arXiv:2308.12833 [cs.CL], https://arxiv.org/abs/2308.12833.

67 Mitchell et al., *supra* note 3, at 7.

68 Richard Ngo et al., *The Alignment Problem From a Deep Learning Perspective*, arXiv, 1, 1–2 & 11, (May 4, 2025), arXiv:2209.00626 [cs.AI], https://arxiv.org/abs/2209.00626.

69 Usman Anwar et al., *Foundational Challenges in Assuring Alignment and Safety of Large Language Models*, arXiv, 1, 34 (Sep. 6, 2024), arXiv:2404.09932 [cs.LG], https://arxiv.org/abs/2404.09932.

70 Mitchell et al., *supra* note 3, at 6.

71 *Id.*

72 Alexander Pan et al., *Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the Machiavelli Benchmark*, *in* 202 Proceedings of the 40th International Conference on Machine Learning Research 1–2 (2023).

73 Chan et al., *supra* note 43, at 960.

74 United Nations, *Governing AI for Humanity: Final Report* 29 (Sept. 2024), https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.

3 Foreseeability in U.S. Negligence Law

In numerous jurisdictions across the world, foreseeability is an essential element of the negligence doctrine in tort law. It exists to ensure that a defendant is only held liable for conduct they were aware of or could foresee.[75] The underlying logic is that a reasonable defendant cannot take "precautionary steps or modify (their conduct)" if they are not aware of a risk.[76] Consequently, the requirement of foreseeability rests on the understanding that liability should only ensue if fault or fault-like behavior can be traced, and awareness of the risk is the first step in determining fault. In this Part, this article will review the role that foreseeability plays in U.S. negligence law. Findings will almost entirely be drawn from judicial doctrine, which is the main source of common law negligence in the country. However, academic work will also be used to make further sense of findings.

3.1 The Origins of Foreseeability in U.S. Negligence Law

U.S. negligence law developed largely from English common law, which originally embraced a strict cause-and-effect model. Under early versions of the writ of trespass, a wrongdoer was liable for the direct consequences of their actions, regardless of whether the harm was reasonably foreseeable.[77] These early rules focused on physical causation rather than mental states or foresight. Legal maxims from the period reflected this approach. For example, Francis Bacon's seventeenth-century principle *in jure non remota causa sed proxima spectator* ("in law, the near cause is heeded, not the remote") expressed a preference for limiting liability to direct, proximate causes.[78] Courts gradually began to grapple with how to distinguish proximate or natural consequences from those deemed remote or extraordinary.

A turning point came in the 1773 case of *Scott v. Shepherd*.[79] In that decision, Justice Blackstone, writing in dissent, arguably introduced an early form of foreseeability reasoning. He suggested that an intervening act would not necessarily break the chain of causation if the ultimate injury was a predictable outcome of the

75 Mulheron, *supra* note 9.

76 *Id.*

77 Daniel Herron et al., *The Evolution of Foreseeability in the Common Law of Tort*, 35 N. E. J. of Legal Stud. 1, 2–4 (2016).

78 Patrick J. Kelley, *Proximate Cause in Negligence Law: History, Theory, and the Present Darkness* 69 Wash. U. L. Rev 49, 54 (1991) citing to Francis Bacon, *A Collection of Some Principal Rules and Maximes of the Common Laws of England, in* The Elements of the Common Laws of English Regula 1, at 1 (1630 & photo reprint 1978).

79 96 Eng. Rep. 525 (K.B. 1773); Herron et al., *supra* note 78, at 7–8.

original wrongful act.[80] While the case still operated within a causation-based framework, it signaled a shift toward moral evaluations of blame that considered the foreseeability of consequences.

By the nineteenth century, the industrial revolution introduced new forms of risk and made the boundaries of liability more difficult to manage. Courts in both England and the U.S. increasingly began to incorporate foreseeability into the evaluation of negligence. Rather than relying solely on mechanical chains of causation, judges required that the harm be recognizable in advance by a reasonable person before imposing liability. Oliver Wendell Holmes Jr. captured this shift in his observation that an act is excusable only if the actor "neither did nor could, with due care, have foreseen any harm resulting."[81] Holmes's formulation helped reshape the inquiry around what a reasonable actor in the defendant's position would have anticipated. Courts in this period began to distinguish between proximate harms, which were reasonably foreseeable, and remote harms, which were not.

An early application of this logic appears in the 1866 case of *Ryan v. New York Central R.R. Co.*[82] There, the court held that a negligent party was liable for the proximate results of a fire it caused, but not for distant and speculative injuries. The court drew a line between "natural and ordinary" consequences of the negligence, which were compensable, and more attenuated harms, which were too remote to support liability.[83] This reasoning reflected a broader doctrinal trend: by the late nineteenth century, foreseeability had become a central organizing principle in negligence law. Courts began to use it not only to define duty but also to evaluate breach and to limit the scope of legal causation. The principle that a person cannot be negligent for failing to prevent a harm that a reasonable person would not have anticipated became firmly embedded in U.S. tort doctrine.[84]

In the early twentieth century, U.S. courts began to apply foreseeability more expansively, particularly in novel situations involving emerging technologies and new social relationships. One landmark example is the 1916 case of *MacPherson v. Buick Motor Co.*,[85] in which Justice Benjamin Cardozo rejected the traditional privity requirement in product liability cases. Cardozo reasoned that a manufacturer could owe a duty of care not only to the immediate purchaser but also to

80 *Id.*

81 Oliver Wendell Holmes Jr, The Common Law, 57 (1881).

82 Ryan v. New York Central R.R. Co. 35 N.Y. 210 (N.Y. 1866).

83 *Id.*

84 Herron et al., *supra* note 78, at 17.

85 MacPherson v. Buick Motor Co. 217 N.Y. 382, 393 111 N.E. 1050 (1916).

any third party whose injury was reasonably foreseeable. He wrote that "injury to others is to be foreseen not merely as a possible, but as an almost inevitable result" of negligence in manufacturing a dangerously defective product.[86] The court's analysis moved foreseeability to the center of the duty inquiry, signaling that foreseeability of harm could independently generate legal responsibility.

At the same time, legal scholars began to formalize foreseeability's doctrinal role. In 1927, Leon Green published *The Rationale of Proximate Cause*, which argued that foreseeability should serve as the principal limitation on liability in negligence and that the question of proximate cause was best left to the jury rather than resolved by rigid legal rules.[87] Green, a leading figure in the legal realism movement, maintained that proximate cause was not a technical test but a policy-laden judgment about fairness, which could be guided but not dictated by precedent. His work provided a theoretical foundation for the idea that foreseeability was a flexible and evaluative tool, not merely a formula for causal analysis.

The concept of foreseeability gained special prominence in U.S. tort law with the 1928 decision in *Palsgraf v. Long Island Railroad Co.*,[88] which remains a foundational case. In *Palsgraf*, a railroad guard accidentally caused a package of fireworks to fall and explode, resulting in an injury to a distant bystander, Helen Palsgraf. The central issue was whether the railroad owed her a duty of care. Justice Cardozo, writing for the majority, held that it did not. He reasoned that duty arises only toward those foreseeably within the zone of danger, and that the harm to Palsgraf was not reasonably predictable at the time of the employee's act.[89] He thus treated foreseeability as a threshold legal question that limits duty. In dissent, Justice Andrews took the opposite view, arguing that duty is owed generally and that foreseeability belongs in the proximate cause analysis, to be resolved by a jury using fairness and policy considerations.[90]

This divergence sparked decades of debate about whether foreseeability functions primarily as a limit on duty or as part of causation. Despite those disagreements, *Palsgraf* solidified foreseeability as a core feature of American negligence law, shaping analysis at every stage – from duty to breach to proximate cause.

---

86 *Id.*

87 Leon Green, The Rationale of Proximate Cause (Vernon Law Book Company 1927) cited in Herron et al., *supra* note 78, at 9.

88 162 N.E. 99 (N.Y. 1928).

89 *Id.* at 99–101.

90 *Id.* at 102–05.

3.2 The Role of Foreseeability in U.S. Negligence Law

U.S. state courts recognize duty of care, breach of duty, factual causation, proximate causation, and injury as the five elements of a *prima facie* negligence case, and specifically examine foreseeability when assessing duty, breach, and proximate cause.[91]

3.2.1 Under Duty of Care

Duty of care analysis in U.S. state courts proceeds on the basis that a defendant cannot be at fault for unreasonable conduct unless the law first recognizes an obligation for the defendant to act with reasonable care toward the plaintiff.[92] Only when a legally cognizable duty exists does it make sense to ask whether the defendant's breach caused harm or whether liability should extend to the resulting injury.[93] The task of establishing duty of care is within the ambit of the courts rather than the fact-finder, as it is a legal question.[94]

There is significant variance in how U.S. courts apply foreseeability when conducting a duty of care analysis. To begin with, and in line with the *Restatement (Third) of Torts,*[95] courts in states like Arizona and Wisconsin have altogether eliminated foreseeability from the duty stage.[96] The rationale for this decision is that the foreseeability question (i) is frequently misused by judges when considered at this stage, and (ii) is meant to be fact-intensive and therefore should be left to juries at the breach and proximate cause stages.[97] However, the vast majority of state courts have not taken this step. Instead, per Jonathan Cardi, "foreseeability is nearly ubiquitous and often cited as the most important factor in duty". Indeed, there are an

91 Dan Dobbs, The law of torts § 114, at 269 (West Group 2000); W. Page Keeton et al., Prosser and Keeton on the law of torts § 30, at 163–64 (5th ed. West Publishing Company 1984); Leon Green, Judge and jury 66 (Vernon law Book Company 1930) all cited in Cardi, *supra* note 8, at 743.

92 Taylor v. Smith, 892 So. 2d 887 (Ala. 2004); Div. of Corr. v. Neakok, 721 P.2d 1121, 1125–1126 (Alaska 1986); William L. Prosser, Palsgraf *Revisited*, 52 Mich. L. Rev. 1, 12 (1953).

93 Cardi, *supra* note 8, at 751.

94 Kirksey v. Tonghai Mar., 5th Cir. 2008, 5; Alani Golanski, *A New Look at Duty in Tort Law: Rehabilitating Foreseeability and Related Themes*, 75 Albany L. Rev. 227, 232 (2011–2012).

95 Restatement (Third) of Torts: Liability for Physical Harm § 7 cmt. j (Am. L. Inst. 2010).

96 Gipson v. Kasey, 150 P.3d 228, 231–32 (Ariz. 2007); Behrendt v. Gulf Underwriters Ins. Co., 768 N.W. 2d 568, 575–76 (Wis. 2009). Other cases include: Simonetta v. Viad Corp., 197 P.3d 127, 131 n.4 (Wash. 2008); Rosengren v. City of Seattle, 205 P.3d 909, 913 (Wash. Ct. App. 2009); Demshick v. Cmty. Hous. Mgmt. Corp., 34 A.D.3d 518, 520 (N.Y. App. Div. 2006); Madden v. Ceglio, 841 N.Y.S.2d 821, 822 (N.Y. Sup. Ct. 2007).

97 Leon Green, *Proximate Cause in Texas Negligence Law*, 28 Tex. L. Rev. 755, 772 (1950); Leon Green, *Are Negligence and Proximate Cause Determinable by the Same Test? – Texas Decisions Analyzed* 1 Tex. L. Rev. (1923) 423, 434–35; Leon Green, *Foreseeability in Negligence Law*, 61 Colum. L. Rev. 1401, 1421 (1961).

exceedingly few number of courts which base their duty findings on factors that exclude foreseeability.[98]

After reviewing case law from all the U.S. jurisdictions, Cardi found that in nearly equal share, some state courts prioritize the foreseeability of the harm in question, others the foreseeability of the harm to the plaintiff, others the foreseeability of the risk and others simply focus on foreseeability in general.[99] There are also substantial differences in how state courts describe the object of the duty-foreseeability assessment. Most conduct a particularized assessment of each plaintiff's injury or risk to decide if there was foreseeability.[100] Others claim that the object of inquiry is meant to be whether the conduct in question created a class of risk that the plaintiff's injury fits into.[101] This is ideally supposed to be a categorical analysis, but it often ends up being case-specific because the court has to decide whether the particularized facts show a fit. Finally, a handful of state courts consistently conduct a genuinely categorical analysis that focuses on whether the defendant's actions created some general range or risk of harm.[102] In this case, the foreseeability of the plaintiff's specific injury is analysed in the proximate cause stage.

At the heart of these differences in approach are questions like: Is duty based on moral obligations or instrumentalist concerns?[103] Is duty relational or act-centered?[104] And if duty is relational, how strongly or weakly relational is it?[105] Implicitly, the answers to these questions guide the decisions that U.S. state courts make regarding duty-foreseeability.

It is worth noting that little has changed since Cardi's comprehensive study was published in 2011. While some more state courts have moved to eliminate foreseeability from their duty analysis,[106] most continue to treat foreseeability as relevant to duty.[107] And in response to the increasingly frequent way juries were *de facto*

98 W. Jonathan Cardi, *The Hidden Legacy of* Palsgraf*: Modern Duty Law in Microcosm*, 91 B. U. L. Rev. 1873, 1884 (2011).
99 *Id.* at 1884–1885.
100 *Id.* at 1886.
101 *Id.*
102 *Id.*
103 Golanski, *supra* note 95, at 235.
104 *Id.*
105 Esper & Keating, *supra* note 11, at 1240–41.
106 *See e.g.*, New Mexico (see Rodriguez v. Del Sol Shopping Center Associates, L.P., 326 P.3d 465 (N.M. 2014) (en banc)) and Kentucky (Kenton v Kentucky Easter Seals Society, Inc., 413 S.W.3d 901 (Ky. 2013)).
107 Jennifer F. Thompson & Deborah Fetra, *New Mexico Supreme Court Eliminates Foreseeability from Tort Duty Analysis*, The Federalist Society State Court Docket Watch (Dec. 11, 2014), https://fedsoc.org/scdw/new-mexico-supreme-court-eliminates-foreseeability-from-tort-dutyanalysis#:~:text=40%20According%20to%20the%20California,Wiener (last visited June 29, 2025).

being left to decide duty-foreeseability, some courts have clarified that duty-foreeseability is a decision for the judge.[108]

3.2.2 Under Breach

A breach analysis generally tries to establish whether the defendant acted to the required standard,[109] and it is usually within the ambit of a jury.[110] Within that analysis, a sub-analysis on foreseeability is done by the jury to decide whether breach occurred. Essentially, once a judge has established that a defendant owed a duty of care and articulated the applicable "standard of care", it is for the jury to determine whether the defendant's conduct fell short of that standard.[111]

In assessing breach-foreseeability, U.S. juries are typically directed to evaluate foreseeability from the standpoint of a reasonable person under the circumstances.[112] This reasonableness review is conducted by considering several factors, including: (i) "The degree of foreseeable likelihood, from the point of view of a reasonable person in defendant's position, that defendant's actions might result in injury",[113] (ii) how severe the foreseeable harms are, and (iii) the precautions that could have been taken.[114] The likelihood and severity of foreseeable harm together define the "risk" posed by a person's actions.[115] Thus, U.S. state courts generally hold that as that risk, whether through increased likelihood or heightened severity of injury, rises, so too does the duty of care the actor must exercise.[116]

When assessing what it means to create a foreseeable risk of harm in a breach analysis, most U.S. state courts do not require that the exact injury the plaintiff suffered be foreseeable.[117] Cardi gives an example of two friends who drink to the same extent, then drive home while intoxicated. One driver crashes into a mailbox, causing only property damage, while the other kills two teenagers by colliding with a parked car. Though the wrongful results differ – and hence the damages differ – both

108 See, for example Vermont's position as taken in Fagnant v Foss, 82 A.3d 570, 576–77 (Vt. 2013).
109 Keeton et al., *supra* note 92, at § 30, 164.
110 Cardi, *supra* note 8, at 744.
111 Smith v. Frank, 207 F. App'x 617, 620 (6th Cir. 2006) citing McClenahan v. Cooley, 806 S.W.2d 767 (Tenn. 1991); Cardi, *supra* note 8, at 744.
112 Cardi, *supra* note 8, at 744–45.
113 *Id.* at 745.
114 United States v. Carroll Towing Co., 159 F.2d 169, 173–74 (2d Cir. 1947); Markowitz v. Arizona Parks Bd., 706 P.2d 364, 369 (Ariz. 1985).
115 Zettle v. Handy Mfg. Co., 998 F.2d 358, 360 (6th Cir. 1993); McKinney v. Louisiana Nat. Bank, 416 So. 2d 948, 951 (La. Ct. App. 1982).
116 Loilar v. Poe, 622 So. 2d 902, 908 (Ala. 1993); Erbrich Prods. Co. v. Wills, 509 N.E.2d 850, 855 (Ind. Ct. App. 1987); Lovell v. Oahe Elec. Coop., 382 N.W.2d 396, 398 (S.D. 1986); Cardi, *supra* note 8, at 745.
117 Dan B. Dobbs, The Law of Torts § 143, at 335 (2000) cited in Cardi, *supra* note 8, at 746.

drivers are equally negligent, because each exposed others to the same spectrum of foreseeable dangers by driving drunk. The breach inquiry focuses on the risk of harm created by the conduct in question, not on the precise injury that ultimately occurred.[118] Furthermore, when conducting such evaluations, U.S. courts focus on what the defendant actually knew or should have known at the time. This aligns with the general view that hindsight bias is to be avoided at all costs.[119]

3.2.3 Under Proximate Cause

Proximate cause is the element under which U.S. state courts examine how closely a defendant's negligent actions are linked to the plaintiff's injury.[120] It obtains its status from the legal principle that, despite the fact that "the consequences of an act go forward to eternity," liability must be confined to prevent unlimited responsibility for every remote result of wrongful conduct.[121] Through proximate cause, liability is estopped where the harm is so far removed from the risks the defendant created that holding them responsible would be unjust or unworkable.[122]

U.S. state courts determine proximate cause by asking whether the injury that happened "was of the same general nature as the foreseeable risk" from the defendant's conduct.[123] Just like the breach-foreseeability analysis, the proximate cause-foreseeability inquiry is left to juries because it is designed to be fact-intensive. As it stands, most of the jurisprudence from U.S. state courts requires those carrying out the inquiry to be attentive to the specific injury of the specific plaintiff.[124] Consequently, even if some harm may appear foreseeable, recovery may fail at proximate cause if the negligence caused (a) an unforeseeable type of injury,[125] (b) harm by an unexpected chain of events,[126] or (c) injury to someone outside the defendant's foreseeable zone of risk.[127]

---

118 *Id.* at 746–47.

119 Jeffs v. West, 2012 UT 11, ¶ 28, 275 P.3d 228, 235 (Utah 2012).

120 Keeton et al., *supra* note 92, at § 41, 264.

121 *Id.*

122 Cardi, *supra* note 8, at 747.

123 Tetro v. Town of Stratford, 458 A.2d 5, 7–8 (Conn. 1983).

124 Cardi, *supra* note 8, at 749.

125 Baltimore City Passenger Ry. Co. v. Kemp, 61 Md. 74, 82–83 (1884).

126 Bunting v. Hogsett, 21 A. 31, 32–33. Here, the court held that as a legal matter that a railroad passenger's injury was foreseeable when a collision derailed an engine, sent it careening around a circular track, and it struck the passenger in a subsequent crash.

127 In re Guardian Cas. Co., 2 N.Y.S.2d 232 (N.Y. App. Div. 1938). The Court held that as a legal matter that a bystander's death was foreseeable when a taxi crashed into a building, dislodged a stone, and that stone later fell and killed the victim during vehicle removal.

4 A Path Forward for Conceptual Reasoning About Foreseeability of Harms Caused by Autonomous AI Agents

As noted in Part II, foreseeability analyses in negligence typically involve both fact-specific and categorical arguments. This Part argues that when properly weighed, certain categorical arguments justify adjustments to negligence-foreseeability doctrine. In particular, it explains why, in cases involving personal injury caused by autonomous agents, such adjustments should favor plaintiffs bringing claims against developers. Many of the rationales commonly invoked in favor of strict liability for harms caused by AI systems can apply here. These include no-fault-based arguments (for example, the importance of guaranteeing compensation for victims[128] and assigning the cost of injury to those who profit from AI[129]) and fault-based arguments (for instance, the moral and structural accountability of developers in light of design choices that contribute to harm[130]).

This Part focuses on a specific, underexplored fault-based argument: that both the black box problem and the technical autonomy of AI systems arise from abstraction choices made by developers. I focus on this argument because scholars and commentators frequently identify the black box problem and technical AI autonomy as reasons to approach questions of foreseeability with caution. After that analysis, the discussion turns to why the personal injury context is especially important in AI harm cases, and concludes by explaining why tort law is a fitting option with which to the pursue the goal of deterrence.

4.1 The Black Box Problem and Technical AI Autonomy

This section will demonstrate how the black box problem and AI autonomy often shape reasoning around liability for AI harms and then turn to proposing how the two should shape such reasoning.

128 David C. Vladeck, *Machines without Principals: Liability Rules and Artificial Intelligence*, 89 Wash. L. R. 117, 120–122, 127 (2014).

129 *Id* at 146–47.

130 Mark Chinen, *The Co-Evolution of Autonomous Machines and Legal Responsibility*, 20 Vand. J. L. & Tech 338, 361–63 & 365–66 (2016); Selbst, Venkatasubramanian & Kumar, *supra* note 12, at 429.

4.1.1 How the Black Box Problem and Technical AI Autonomy Shape Reasoning Around Accountability for AI Harms

The black box problem and AI autonomy dominate scholarly debates about legal accountability for AI harms. Both issues are usually framed as inconveniences that compel us to rethink our typical use of legal doctrines designed for determining liability. Writing about the topic, Andrew Selbst and et al. have correctly observed that:

> black boxes are ubiquitous in narratives surrounding accountability for technological harms, especially in those involving machine learning and artificial intelligence. The black box is usually the villain of the story, the reason that there cannot be accountability for the harms.[131]

Let us consider some examples. Yavar Bathaee has argued that legal doctrines such as causation and foreseeability might prove inadequate in addressing conduct of AI models because of the black box problem. His argument can be summarized as follows: AI models can uncover hidden patterns in massive datasets, devise unintuitive solutions, act at speeds beyond human capability, or base decisions on complex, high-dimensional relationships that no human can visualize.[132] Thus, the question becomes, if the very individuals who built the system cannot foresee its behavior or results, is it possible to expect a "reasonable person" to do so?[133]

This pattern of reasoning is echoed by a wide group of scholars including Ryan Calo,[134] Mathew Scherer,[135] Weston Kowert,[136] Kenneth Abraham,[137] Robert Rabin,[138] Yaniv Benhamou and Justine Ferland.[139] It is particularly useful to highlight Benhamou and Ferland's contribution, which argues that determining the root cause of an AI system's malfunction is essential for proving a breach of duty in tort claims. Yet, when a plaintiff cannot retrace the AI's data-processing steps and reconstruct its

131 Selbst, Venkatasubramanian & Kumar, *supra* note 12, at 416.

132 Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 Harv. J. of L. and Tech. 889, 897–905 (2018).

133 *Id.* at 924.

134 Ryan Calo, *Robotics and the Lessons of Cyberlaw* 103 Cal. L. Rev. 513, 555 (2015).

135 Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, And Strategies*, 29 Harv. J. of L. and Tech. 353, 365 (2016).

136 Weston Kowert, *The Foreseeability of Human Artificial Intelligence Interactions* 96 Tex. L. Rev. 181, 183 (2017).

137 Kenneth S. Abraham and Robert L. Rabin, *Automated Vehicles and Manufacturer Responsibility for Accidents: A New Legal Regime for a New Era*, 105 Va. L. R. 127, 134 (2019).

138 *Id.*

139 Benhamou & Ferland, *supra* note 56, at 8.

decision-making, they cannot meet the basic evidentiary requirements for fault or causation, and their claim will likely fail.[140]

Closely connected to the black box problem, technical AI autonomy is the other idea that plays an outsize role in scholarly debates about legal accountability for harms caused by AI. Some examples follow.

Scholars like Benhamou,[141] Ferland[142] and Chinen[143] have argued that as AI systems gain greater autonomy, assigning liability becomes increasingly fraught because no single human actor can reliably anticipate or control the machine's decisions.[144] Benhamou and Ferland claim that traditional fault-based frameworks – such as foreseeability of harm – work only when one can trace the agent's actions back to a specific person (such as a manufacturer or user) who could have foreseen and prevented the damage.[145] When AI learns and adapts on its own, interpreting data and refining its behavior without human input, its outcomes lie beyond anyone's reasonable foresight. Moreover, scholars like Chinen argue that in such cases, it is virtually impossible to identify whose breach of duty or fault caused the injury, since the system's autonomous evolution severs the link between any human decision and the harm done. Consequently, foreseeability cannot be proved, leaving traditional liability rules insufficient for attributing responsibility.[146]

Benhamou and Ferland have also suggested that when applying tort law to AI-driven conduct, the inherent unpredictability of autonomous systems undermines both the assessment of breach and the establishment of causation.[147] Because no one can reliably forecast all the ways an AI might cause harm, neither operators nor developers can know which precautions, tests, or safety measures would suffice to prevent injury.[148] As a result, it is often unrealistic to expect human stakeholders to guard against all potential risks, and even diligent development, training, and quality-control practices may not be deemed negligent if an AI later behaves unexpectedly. Furthermore, they argue that even where a human error in handling the

140 *Id.*

141 *Id.* at 6–7.

142 *Id.*

143 Chinen, *supra* note 131, at 360.

144 Benhamou & Ferland, *supra* note 56, at 6–7; Chinen *supra* note 131, at 360.

145 Benhamou & Ferland, *supra* note 56, at 6–7.

146 Chinen, *supra* note 124, at 359–60.

147 Benhamou & Ferland, *supra* note 56, at 7.

148 *Id.* at 9. *See also*, Omri Rachum-Twaig, *Whose Robot Is It Anyway?: Liability for Artificial-Intelligence-Based Robots*, Univ. Ill. L. Rev. 1141, 1164 (2020); Woodrow Barfield, *Liability for Autonomous and Artificially Intelligent Robots*, 9 Paladyn J. of Behavioral Robotics 193, 200 (2018).

AI can be shown, the AI's unforeseeable actions frequently sever the causal link between that error and the victim's harm, leaving traditional tort law principles deficient.[149]

Mark Chinen contends that the existing tension between distributed development and the need for foreseeability undermines both moral blameworthiness and legal fault, and thus calls for a fundamental rethinking of how we assign liability for truly autonomous models.[150] There is certainly some truth to the fact that the black box problem and AI autonomy significantly complicate any negligence – and indeed legal – tests designed to conclude whether there is fault in a defendant's conduct. It is also true that the two phenomena must be grappled with in any legitimate process for determining legal fault. However, as this article will show in the next section, by centering these two phenomena in the manner described, current AI accountability debates have sidelined an important perspective.

4.1.2 The Black Box Problem and Technical AI Autonomy as Products of Abstraction Choices

Formal abstraction lies at the heart of computer science and system design. It means specifying a component using only its inputs, outputs and the functional relationship between them.[151] When formal abstraction takes place, abstraction boundaries are created, and some features and issues are "abstracted away". Selbst et al. use an example of a coffee machine to make this point. The inputs are water and coffee beans, and the output is coffee. The engineer's definition of the problem does not include where the water will come from or where the beans will be sourced from,[152] meaning that the abstraction boundary leaves out those questions.

Engineers draw abstraction boundaries implicitly through the technical design decisions that they make. As they make these decisions, they also implicitly make claims about what the object is and where the lines of responsibility fall.[153] In most cases, their choices are motivated by factors such as efficiency, economic and marketing concerns, and existing legal requirements.[154] There is evidence that, when the situation calls for it, computer scientists choose their abstraction boundaries

149 Benhamou & Ferland, *supra* note 56, 9.

150 Mark Chinen, Law and Autonomous Machines: The evolution of autonomous machines and legal responsibility, 66–71 & 73 (Edward Elgar Publishing 2019).

151 Selbst, Venkatasubramanian & Kumar, *supra* note 12, at 424–429.

152 *Id.* at 426.

153 *Id.* at 429.

154 *Id.*

broadly enough to encompass a range of factors that would typically be considered "non-technical".[155]

When AI developers build deep learning models on architectures like transformers, the unexplainability and unpredictability that carries through is not an inherent property of the models but a product of the developers' technical decisions. In other words, the unexplainability and unpredictability arises from formal abstraction. Architectures like transformers are extremely complex, high-dimensional and non-linear, with billions of parameters interacting dynamically.[156] Developers adopt these kind of complex architectures because they tend to yield higher accuracy or enable functionalities that simpler, explainable models cannot easily achieve.[157] In doing so, they are making a trade-off insofar as they abstract the problem in purely mathematical terms (letting the algorithm itself derive whatever features or rules maximize accuracy) and sacrifice interpretability in the process.

We therefore have black box models because AI developers choose to prioritize some goals over others. In place of transformers, for example, developers could opt for more transparent models like simple decision trees, rule-based systems, attention-based models (without full transformer complexity) or shallow neural networks.[158] They often decide against these architectures because their adoption would result in models that are less accurate or scalable. The important point is that opacity is a consequence of developers' decision-choices. Indeed, the widespread adoption of black-box models in the industry is evidence that explainability has not been considered a make-or-break priority during the design phase.

The same is true of the technical autonomy of AI. As shown in the preceding section, many legal scholars discuss technical autonomy as if it is an intrinsic feature of deep learning models. However, just like the black box problem, it is a result of abstraction choices. Deep learning models behave autonomously because AI

---

155 Jing Nathan Yan et al., *Fairness Practices in Industry: A Case Study in Machine Learning Teams Building Recommender Systems*, arXiv, 1, 1 (May 26, 2025) arXiv 2505.19441 [cs.HC], https://arxiv.org/abs/2505.19441; Kenneth Holstein et al., *Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?*, arXiv, 1, 1 (Jan. 7, 2019), arXiv:1812.05239 [cs.HC], https://arxiv.org/abs/1812.05239.

156 Hendrycks, *supra* note 65, at 91–92; Ashish Vaswani, et al., *Attention is all you need*, in Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000, 6000 (2017).

157 Yi Tay et al., *Efficient Transformers: A Survey*, arXiv, 1, 2 (Sept. 14, 2020), arXiv:2009.06732 [cs.LG], https://arxiv.org/abs/2009.06732.

158 Barnaby Crook, Maximilian Schlüter & Timo Speith, *Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)*, arXiv, 1, 2 (Sept. 26, 2023) arXiv:2307.14239 [cs.AI], https://arxiv.org/abs/2307.14239; André Assis, Jamilson Dantas & Ermeson C. Andrade, *The Performance-Interpretability Trade-Off: A Comparative Study of Machine Learning Models*, 11 J. Reliable Intelligent Envs. 1, 1 (2025).

developers design them to operate with a degree of independence and deliberately avoid hard-coding every decision. At the design phase, engineers decide to relinquish control over certain decisions and instead encode objectives or learning rules that the model will follow. When reinforcement learning is used, for example, developers specify a reward function and then allow the model to explore strategies to maximize that reward.[159] The model's technical autonomy is a consequence of the developers' decision to use reinforcement learning, and the trade-off is an inability to specify how the model will handle every possible situation.

There are three strong counterarguments that could be offered as rejoinders to the argument staked out above. The first could be that abstraction may obscure certain features, but it does not generate unpredictability or opacity in any meaningful sense. Rather, unpredictability and opacity are properties that result from intrinsic factors such as the lack of AI embodiment, emergent behavior, and epistemic limitations during design.[160] Although it is persuasive, this argument only partially responds to the point, which is that the engineering of deep learning models is not categorically distinct from abstraction; rather, it is the natural continuation of abstraction into domains of extremely high complexity and probabilistic behavior. If the abstraction choices were to be less technical and more reflexive, that could significantly reduce AI opacity and unpredictability. The failure to do more reflexive abstraction is a choice, especially when known risks are ignored.

The second counterargument could be that the substantial investment by AI companies in model interpretability research[161] is evidence that interpretability and predictability have not been abstracted away, as these companies are working to make their models more interpretable, understandable and steerable. However, this argument ignores the fact that investment in interpretability often reflects a reactive effort to mitigate the challenges posed by the complexity and opacity of deep learning models. Indeed, interpretability is often treated as a secondary concern to performance, with model developers focusing primarily on accuracy and efficiency while interpretability is worked on as a secondary issue. Thus, the current state of interpretability research only reinforces the continuing abstraction of deep learning systems.

---

159 Richard S Sutton and Andrew G Barto, Reinforcement learning: An introduction, 57–8 (MIT Press, 2018).

160 Mihaly Heder, *The Epistemic Opacity of Autonomous Systems and the Ethical Consequences, AI and Society*, 38 AI & Soc 1819, 1821–26 (2023).

161 Diogo Carvalho et al., *Machine Learning Interpretability: A Survey on Methods and Metrics* 8(8) Electronics 1, 4–5 (2019); Sungsoo Ray Hong, et al., *Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs in* Proc. ACM Hum. Comput. Interact. 4, Article 68, 1–26 (May 2020); Google LLC, *Responsible AI Progress Report*, 14 (Feb. 2025), https://ai.google/static/documents/ai-responsibility-update-published-february-2025.pdf.

The third counterargument could be that the ubiquity of deep neural networks, transformer architectures, reinforcement learning et cetera signals not irresponsibility but shared industry consensus about what counts as appropriate and efficient design. In other words, the fact that engineers across companies and research institutions converge on these methods suggests that the methods are not only acceptable but reasonable within the professional standard of care. This response is weak since widespread adoption does not usually equate to normative legitimacy. As many tort law courts have recognized, industry custom can still fall short of the expected standard of care.[162]

As has been shown in section A (1) in this Part of the article, many legal scholars have concluded that the existence of the black box problem and AI's technical autonomy have made it unusually difficult to find fault in AI developers' and deployers' decisions when the models they build result in harm. On account of the abstraction argument, this article contends that this frame of thinking should be reconsidered. Once we see the black box problem and AI's technical autonomy as the automatic results of abstraction choices made by developers, we will not be so quick to free them of responsibility. In the words of Selbst et al., courts should adopt approaches that show they understand that "the technological arrangements themselves represent substantive claims that bear directly on legal proceedings."[163]

When taken seriously, the argument that the black box problem and AI autonomy are a result of abstraction invites a reexamination of legal standards governing accountability for AI harms. One such standard is foreseeability. I argue that the abstraction choices which result in black box models that are technically autonomous (hence opaque and unpredictable) justify a corresponding relaxation of foreseeability tests in favor of plaintiffs. Neither the black box problem nor technical AI autonomy should be viewed as exculpatory since both are the product of industry design choices that prioritize accuracy, efficiency and commercial success over interpretability and predictability. When such prioritization leads to models which cause harm, the legal system has reason to intervene on behalf of the injured party. Yet specific doctrinal frameworks can frustrate that possibility. Foreseeability, in particular, sometimes operates in ways that frustrate plaintiffs and constrain judicial recognition of novel risks. Where foreseeability doctrine serves to insulate AI developers from accountability in the face of design-driven opacity and autonomy, it ought to be relaxed. The precise contours of such a doctrinal shift merit close analysis. I take up that question in the next Part. First, however, some additional observations are warranted.

---

162 Klimko v. Rose, 84 N.J. 496, 506 n.4, 422 A.2d 418 (1980); Trimarco v. Klein, 56 N.Y.2d 98, 436 N.E.2d 502 (N.Y. 1982).

163 Selbst, Venkatasubramanian & Kumar, *supra* note 12, at 417, 420.

4.2 The Weight of Personal Injury in Tort Law

In tort law, the integrity of an individual's mind and body has long been regarded as a fundamental interest, one that justifies the subordination of other weighty societal goals. As Christian Witting has observed, there is broad consensus that tort law affords higher protection to interests in bodily integrity, property, and reputation compared to purely financial interests.[164] In a similar vein, Dilan Esper and Gregory Keating argue that twentieth-century tort law "got something important right": namely, the conviction that the physical integrity of persons is a more pressing legal interest than either absolute control over real property or unrestrained economic liberty.[165] Judges also seem to be aligned with this position. For example, U.K. courts frequently use legal standards that are claimant-friendly when deciding remoteness-foreseeability in cases where personal injury has been suffered.[166]

Witting offers two arguments why tort law's prioritization of personal injury over financial interests is correct. The first is that interests in the body, property and reputation are personal and irredeemable while the cost of financial loss for corporate defendants is merely economic and calculable.[167] The second centers on the long-term costs of personal injury to a plaintiff. In this argument, affliction of personal injury is understood as a "double blow" to plaintiffs, potentially creating challenges for both them and their dependents. On the other hand, shareholders are better placed to cope with financial losses.[168]

In a similar vein, Gregory Keating argues that physical harms are special because they disable agency,[169] and that tort law should impose duties to avert physical harm to the greatest extent reasonably possible rather than defer to utilitarian cost-benefit analysis.[170] Keating further argues that American health, safety, and environmental law often requires more than efficient precaution against physical harm, and that this is justified because avoidance of harm has priority over provision of benefit, a priority rooted in the value of individual autonomy.[171] These

164 Christian Witting, Liability of Corporate Groups and Networks 288–89 (Cambridge Univ. Press 2018).
165 Esper & Keating, supra note 11, at 1239.
166 Mulheron, *supra* note 9 at 491; Hughes v Lord Advocate [1963] A.C. 837, 1 All E.R. 705, 2 W.L.R. 779 (H.L.); Jolley v Sutton LBC [2000] 1 W.L.R. 1082, 1091 (H.L.); Lear v Hickstead [2016] EWHC 528 ¶ 99 (Q.B.); ST v Maidstone and Tunbridge Wells NHS Trust [2015] EWHC 51 (Q.B.).
167 Witting, *supra* note 165, at 289.
168 *Id.*
169 Gregory C. Keating, *Is Cost-Benefit Analysis the Only Game in Town?*, 91 S. Cal. L. Rev. 195, 217–30 (2018).
170 *Id.* at 230–33.
171 Gregory C. Keating, *Principles of Risk Imposition and the Priority of Avoiding Harm*, 36 Revus: J. Const. Theory & Phil. L. 1, 5 (2018).

insights support the claim that tort law's solicitude for personal injury is not a contingent doctrinal preference but a principled commitment, one that should inform how courts approach foreseeability in cases where autonomous agents cause personal injury.

Although bodily integrity and autonomy are generally considered especially important or "thick" per William Lucy,[172] their normative importance seems to vary across different torts. This observation may seem puzzling. How should we make sense of the fact that in torts like trespass to the person, even minimal interferance with bodily integrity and autonomy triggers liability while in torts like negligence, similar interferance must be accompanied by additional elements for liability to be triggered? Does this suggest that bodily integrity and autonomy are less important in the negligence framework? As Lucy correctly observes, the explanation lies less in any principled moral distinction and more in the historically contingent evolution of tort doctrine.[173] In particular, the development of the common law action on the case was non-linear and piecemeal,[174] making it difficult to extract a fully coherent moral narrative from its doctrinal architecture.

4.3 Tort Law's Deterrent Effect

As Part I of this article has shown, autonomous agents are especially risk laden. Their high levels of capability and technical autonomy could render them capable of causing harm on an unusually large scale. Autonomous agents' risks of misuse and misalignment underscore the need for a legal framework that discourages AI developers from designing and deploying such systems without the utmost caution. Tort law is well-suited to serve this deterrent function since one of its core aims is to reduce socially harmful conduct by incentivizing actors to avoid excessive or unjustified risk.[175] Moreover, tort law can be particularly effective in regulating corporate behavior. As Christian Witting has demonstrated, organizations are institutionally well-positioned to comprehend, internalize, and respond to legal directives because their structural features enable them to process and implement the behavioral signals that tort law sends.[176] Given that the leading developers of advanced AI are corporations, there is strong reason to believe that tort decisions can influence them to pursue more cautious and responsible development practices.

---

172 William Lucy, Philosophy of Private Law 218 (Oxford Univ. Press 2007).
173 *Id.*
174 *Id.*
175 Witting, *supra* note 165, at 347.
176 *Id.* at 349–351.

Taken together, the foregoing analysis supports a targeted relaxation of the foreseeability requirement in negligence, favoring plaintiffs in cases involving personal injury caused by autonomous agents and brought against AI developers. This conclusion is further supported by the fact that U.S. courts have found in some cases that it is fair to shift the burden of proof to a defendant where their own conduct created evidentiary barriers for the plaintiff.[177] This article advocates for the same logic to be accepted here.

Part IV examines the appropriate contours of such a modification. The analysis rests on three guiding principles:

(i) Where existing foreseeability doctrine enables plaintiffs to establish liability without undue burden, it should remain intact;
(ii) Where the doctrine imposes categorical obstacles that unduly limit plaintiffs' ability to establish foreseeability, it should be recalibrated to lower those thresholds; and
(iii) Any reforms must preserve a role for fact-intensive assessment in determining liability in individual cases.

The next Part of the article sets out the specific doctrinal reforms proposed.

5 Doctrinal Implications for Foreseeability in U.S. Negligence Law

In Part III, this article argued that legal standards governing foreseeability should be relaxed in favor of plaintiffs in cases where autonomous agents cause personal injury, and an AI developer is named as the defendant. This Part of the article advances that argument by setting out specific recommendations for how those standards should be modified. The goal is to develop a framework that modestly shifts the balance toward plaintiffs while preserving a meaningful opportunity for defendants to demonstrate that the harm was not reasonably foreseeable.

The recommendations made here are broadly consistent with the view advanced by Weil, who argues that courts should adopt a more expansive conception of foreseeability.[178] Under this view, liability should turn not on whether the specific harm was predictable, but on whether the general type of risk was reasonably anticipated.[179] Building on this general orientation, the following analysis offers a set

177 Haft v. Lone Palm Hotel, 478 P.2d 465, 474–76 & n.19 (1970); Summers v. Tice, 33 Cal. 2d 80, 83–86, 199 P.2d 1, 3–4 (1948).
178 Weil, *supra* note 13, at 49–50.
179 *Id.*

of doctrinal recommendations tailored to the unique challenges posed by autonomous agents.

5.1 Breach-Foreseeability and Proximate Cause-Foreseeability

In U.S. state courts, decisions regarding breach-foreseeability and proximate cause-foreseeability are generally left to juries, as discussed in Part II of this article. Toward the end of Part III, this article argued that any relaxation of foreseeability standards should still preserve the opportunity for AI developers to demonstrate that the specific personal injury in question was not foreseeable. To ensure this, it is important that the legal doctrine continues to allow for a fact-intensive analysis at some stage of the proceedings. Such an approach is vital because particularized evidence can shape an outcome, even after an initial assumption is made.

Moreover, breach and proximate cause are predominantly questions of fact resolved by juries through case-specific evaluation of the evidence.[180] They are designed to be flexible and responsive to the circumstances of each case, and that design is well suited to the complexity of AI-related harms.[181] If foreseeability standards are to be recalibrated in cases involving autonomous agents, the adjustment should come at a stage of the negligence framework that involves a question of law, not at stages whose fact-intensive character is one of their strengths.

Based on this reasoning, this article argues that the current legal framework surrounding breach-foreseeability and proximate cause-foreseeability in U.S. state courts does not require further adjustment in cases involving personal injury resulting from autonomous agents. The existing framework already facilitates a thorough, fact-based inquiry, which is essential for applying the reasonable person standard in complex cases. Juries, therefore, are well-equipped to consider factors such as whether the AI model was used appropriately and whether it was fine-tuned to a degree that might absolve the AI developer from liability.

While there are certainly difficult questions regarding the jury's understanding of advanced technologies like deep learning, it is worth noting that juries have historically been tasked with determining foreseeability in cases involving complex, scientific matters.[182] Thus, although AI-related cases present new technological nuances, they do not represent a fundamentally different challenge from other cases where juries have dealt with intricate scientific concepts. The understanding of

180 Keeton et al., *supra* note 92, at § 37, 235; Thomas C. Galligan, Jr., *The Structure of Torts*, 46 Fla. St. U. L. Rev. 485, 509 (2019).

181 Galligan, *supra* note 181, at 505, 509–10.

182 Ferebee v. Chevron Chem. Co., 736 F.2d 1529 (D.C. Cir. 1984).

juries may need to be refined, and procedural rules may need to be re-examined, but the current legal doctrine remains largely fit for purpose. Indeed, juries typically do not have an expert understanding of the technologies they decide over,[183] and the standard of proof in civil cases does not require complete certainty.[184]

Jury practices provide further evidence supporting my argument that breach-foreseeability and proximate cause-foreseeability are already well-suited to the task. As demonstrated in Part III, U.S. juries apply breach-foreseeability and proximate cause-foreseeability in a way that gives both plaintiffs and defendants a fair chance at success. To recall, when assessing whether a defendant created a foreseeable risk of harm in breach analyses, no U.S. state court requires that the exact injury suffered by the plaintiff be foreseeable. Similarly, the breach-foreseeability inquiry focuses on the risk of harm created by the defendant's conduct, rather than the precise injury that ultimately occurred. Regarding proximate cause-foreseeability, plaintiffs may fail to prove their case if the negligence resulted in (a) an unforeseeable *type* of injury, (b) harm due to an unexpected chain of events, or (c) injury to someone outside the defendant's foreseeable *zone of risk*. All these do not seem to weigh against a plaintiff, which is why this article argues that no adjustments are needed to the legal standards of breach-foreseeability and proximate cause-foreseeability under U.S. law.

Although the abstraction-choices argument developed in Part III has implications that extend into how juries reason about proximate cause, the existing framework at these stages does not present the same structural barriers to plaintiffs that the duty stage does. The main case for doctrinal reform is therefore made under duty-foreseeability.

5.2 Duty-Foreseeability

We now turn to duty-foreseeability, which this article argues requires adjustment considering the challenges posed by autonomous agents. Before outlining the details of that proposed adjustment, it is useful to revisit the three principal approaches that U.S. state courts have taken in assessing duty-foreseeability, as discussed in Part II.

The first is the strong default duty approach, under which foreseeability is effectively presumed in all cases, and courts do not engage in further inquiry grounded in policy or principle. This functional presumption of duty has been adopted by courts in states such as Oregon. The second is the Third Restatement

183 *Id.*

184 Andrew W. Yung, *Book Review, Phantom Risk: Scientific Inference and the Law*, 7 Harv. J.L. & Tech. 223, 230 (1993); Ferebee v. Chevron Chem. Co., 736 F.2d 1529 (D.C. Cir. 1984).

approach, which similarly begins with a presumption of duty, and effectively foreseeability, but permits that presumption to be displaced by compelling policy or principle considerations. Courts in jurisdictions such as Iowa and Nebraska have followed this model. The third approach is a fact-specific and highly particularized model, in which the existence of duty-foreseeability is determined not by default rules or presumptions, but by close attention to the circumstances of each case. Courts in states such as New York apply this approach, weighing factors such as the relationship between the parties, the nature of the alleged harm, and the context in which the risk arose to determine whether a duty should be recognized.

5.2.1 Proposed Reforms

5.2.1.1 First Approach to Duty-Foreseeability

U.S. courts that follow the first approach to duty-foreseeability do not require any further modification. This model functionally presumes that foreseeability is always satisfied at the duty stage, making it highly favorable to plaintiffs. It therefore aligns with the normative principles outlined at the end of Part III and is well-suited to address cases involving personal injury caused by autonomous agents.

5.2.1.2 Second Approach to Duty-Foreseeability

In light of the argument developed in Part III, this article recommends that courts following the second approach to duty-foreseeability, which begins with a presumption of duty but allows policy considerations to defeat it, should adopt a high threshold for when such considerations may be used to negate foreseeability. According to Cardi's analysis, courts in at least 36 U.S. jurisdictions explicitly identify public policy as a central factor in determining whether a duty exists.[185] Although courts may have valid policy reasons for limiting liability, those concerns are more appropriately addressed at later stages of the negligence framework, such as breach or proximate cause. Addressing policy at those stages allows courts to engage with case-specific facts without prematurely dismissing claims that may be otherwise well-founded.

Based on the arguments proffered in Part III, it would be an injustice for plaintiffs to be barred from recovery at the duty stage. Implementing a high threshold for policy-based denials of foreseeability would preserve courts' flexibility to consider exceptional circumstances while strengthening protection for individuals harmed by autonomous agents. Although it is difficult to define with precision the circumstances that would meet this heightened threshold, it is reasonable to reserve such denials for truly extraordinary contexts. These might include

185 Cardi, *supra* note 8, at 1887.

situations implicating significant national security concerns or cases in which recognizing a duty would create substantial risks to the bodily safety of others. By limiting the use of policy exceptions in this manner, courts can strike a more appropriate balance between systemic considerations and the need to afford meaningful remedies to those injured by autonomous agents.

This position finds support in broader tort theory. Goldberg and Zipursky have argued that a duty of care is ordinarily owed whenever a breach could cause harm to the class of persons like the plaintiff.[186] According to their relational rights account,[187] many cases that courts characterize as involving "no duty" on foreseeability grounds should instead be understood as cases where duty existed but no breach occurred.[188] This view reinforces this article's argument that courts following the second approach should resist the temptation to use policy considerations to negate duty-foreseeability.

One important critique of removing foreseeability from the duty analysis has been advanced by Esper and Keating. They argue that imposing obligations to prevent injuries without regard to foreseeability undermines fundamental principles of legality and moral responsibility.[189] On their view, such obligations go beyond requiring individuals to consider the legitimate claims of others and instead demand a form of conduct that is unreasonably burdensome.[190] Esper and Keating distinguish this from strict liability by noting that negligence liability for unforeseeable harms is unlikely to generate additional precaution, since the harms at issue are, by definition, unforeseeable.[191]

Although this concern raises a legitimate point about the normative limits of negligence, both claims appear overstated. Courts adopting the first and second approaches to duty-foreseeability have not eliminated foreseeability from the negligence framework entirely. They have instead relocated it to the breach and proximate cause stages, where it continues to play a central role. It is therefore inaccurate to characterize this as negligence liability without foreseeability. The second argument (that removing foreseeability from the duty analysis will not promote precaution because actors cannot avoid what they cannot anticipate) is more persuasive but still incomplete. In the context of AI-related harms, precaution is not the only relevant normative goal. Lowering the threshold at the duty stage can serve other important purposes, such as facilitating access to legal protection for injured plaintiffs. Moreover, the removal of foreseeability from the duty inquiry may

186 John Goldberg and Benjamin Zipursky, Recognizing Wrongs 3 (2020).

187 John Goldberg and Benjamin Zipursky, *Torts as Wrongs*, 88 Tex. L. R. 917, 920 (2010).

188 Goldberg and Zipursky, *supra* note 185 at 71.

189 Esper & Keating, *supra* note 11, at 1233–1234.

190 *Id.*

191 *Id.*

encourage general precautionary behavior, such as the use of more transparent (white-box) models[192] or the adoption of safer abstraction strategies in system design.

5.2.1.3 Third Approach to Duty-Foreseeability

Finally, this article argues that U.S. state courts adhering to the third approach to duty-foreseeability should adopt more substantial reforms. As a preliminary matter, it is important to clarify how this approach places plaintiffs at a disadvantage. Unlike categorical or presumptive models, the fact-specific and highly particularized nature of the third approach requires plaintiffs to establish foreseeability at the outset through detailed case-specific evidence. This increases the burden on claimants and heightens the risk of early dismissal, even in cases involving plausible claims of injury. The structure of this approach effectively shifts foreseeability from a guiding legal principle to a factual hurdle, prematurely narrowing access to adjudication. To emphasize this point, it is notable that the key reason why Kentucky courts moved their doctrine away from this high-particularized approach and towards a more categorical approach[193] was to limit judges' ability to grant summary judgement (which they had been overdoing in favor of respondents).[194]

More broadly, as Dilan Esper and Gregory Keating have argued, excessive particularity in negligence analysis undermines core institutional roles. It reflects a failure of courts to articulate coherent legal standards,[195] while simultaneously encroaching upon the jury's traditional responsibility to evaluate facts and apply community norms.[196] It also risks eroding the foundational premise that negligence is an abstract wrong rooted in the failure to take appropriate care, not merely in the technical specifics of a given factual scenario.[197] Addressing these concerns requires recalibrating the third approach to better protect plaintiffs in emerging areas of tort law, particularly in cases involving personal injury caused by autonomous agents.

The most appropriate recalibration of duty-foreseeability in cases involving personal injury caused by autonomous agents is a categorical approach. In

192 Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, arXiv, 1, 1 (Sep. 22, 2019), arXiv:1811.10154v3 [stat.ML], https://arxiv.org/pdf/1811.10154.

193 Shelton v Kentucky Easter Seals Soc'y, Inc., 413 S.W.3d 901 (Ky. 2013).

194 Tia J. Combs & Lucas Harrison, *Minority Report: Kentucky's Evolving Law of Foreseeability*, FM G & G (May 9, 2023), https://www.fmglaw.com/business-litigation/minority-report-kentuckys-evolving-law-of-foreseeability/#:~:text=landowners%20could%20not%20count%20on,much%20harder%20to%20come%20by [https://perma.cc/NF89-CRPK](last visited June 29, 2025).

195 Esper & Keating, *supra* note 11, at 1240.

196 *Id.*

197 *Id.* at 1241.

particular, courts should recognize a baseline category of foreseeable harm that includes any personal injury suffered by any plaintiff as a result of the misuse, misalignment, or malfunction of an autonomous agent developed by the defendant AI company. This formulation reflects the generality required for duty determinations and aligns with the view advanced by Esper and Keating, who argue that legal rules at the duty stage "must be general enough to guide future conduct in similar situations."[198] It also draws normative support from the arguments developed in Part III of this article. While this formulation represents a modest baseline, courts may consider adopting even broader categories as the technology and jurisprudence evolve.

To illustrate how the proposed categorical approach would operate in practice, consider the hospital hypothetical discussed in the introduction of this article. A hospital relies on an autonomous AI agent to schedule patient treatments and assign operating rooms, and the agent erroneously deprioritizes a critical cardiac patient for surgery, resulting in the patient's death.

Under the second approach to duty-foreseeability, the court would begin with a presumption that the developer owed a duty of care. However, the developer could seek to displace the presumption by invoking policy considerations. It might argue, for example, that the social value of AI-assisted healthcare outweighs the risks of imposing open-ended obligations on developers at the duty stage. Under the current framework, such arguments could conceivably succeed. Under the reform proposed here, these arguments fail because they do not rise to the level of truly extraordinary circumstances such as national security risks. The presumption of duty-foreseeability would therefore hold, and plaintiff's claim would proceed.

Under the third, highly particularized approach to duty-foreseeability, the plaintiff's estate would need to demonstrate at the duty stage that this specific scheduling error was a consequence of the developer's conduct. The developer would argue that its agent was designed for efficiency in hospital logistics, that the specific de-prioritization reflected an emergent interaction between the agent's planning algorithms and the data environment of the hospital, and that no reasonable developer could have anticipated this precise failure mode. In the U.S. jurisdictions applying this approach, such arguments could succeed at summary judgement, foreclosing the plaintiff's claim before it reaches a jury.

If the categorical approach proposed here is applied, the analysis would proceed differently. The court would ask whether the plaintiff's injury falls within a broad category, for example misuse or misalignment, and whether there was personal injury. This means the plaintiff in the hypothetical would clear the duty-foreseeability threshold without needing to show that the specific scheduling error

198 *Id.* at 1277.

was predictable. The question of whether the developer acted reasonably would then be resolved under the other sub-tests of negligence.

One might ask whether the domain of foreseeable harm recognized at the duty stage should be further delineated, for example by limiting it to harms that can be traced to a specific abstraction choice. This article resists that approach. As demonstrated in Part III, the abstraction choices that produce opacity and unpredictability are not incidental features of autonomous agents – they are the reason the technology functions as it does. In that sense, the risks created by abstraction are inseparable from the benefits.

Accordingly, this article proposes that the duty recognized at the categorical stage should extend to all personal injuries caused by autonomous agents developed by the defendant, without requiring plaintiffs to trace the harm to a particular abstraction choice. This scope is bounded by two constraints that prevent it from becoming unworkably expansive. The first is the limitation to personal injury, which excludes purely economic losses and confines the duty to the category of harm that tort law has long treated as deserving of the highest protection. The second is the preservation of fact-intensive analysis at the breach and proximate cause stages. A finding of duty would not guarantee liability. Defendants would retain the opportunity to demonstrate that they exercised reasonable care in designing, testing, and deploying the agent, and that the specific harm was not a foreseeable consequence of their conduct.

It is worth acknowledging that the abstraction-choices argument put forth in Part III has implications that extend beyond the duty inquiry. The argument that developers who create opaque and unpredictable systems should not invoke that opacity to defeat a plaintiff's claim speaks not only to whether a duty exists but also to how broadly the scope of the developer's responsibility should extend. In that sense, the argument touches on questions that courts and scholars would ordinarily associate with proximate cause. Nevertheless, this article locates its primary doctrinal intervention at the duty stage for two reasons.

First, the duty stage is where the most significant structural obstacles to plaintiff recovery currently exist. In jurisdictions employing highly particularized duty-foreseeability analyses, claims are often dismissed before they ever reach a jury, foreclosing any possibility of inquiry into breach or proximate cause. Reforming the duty stage is therefore the most urgent priority. Second, proximate cause is predominantly a question of fact resolved by juries through case-specific evaluation of the evidence. Categorical doctrinal reform of the kind proposed in this article is structurally suited to the duty stage, which involves a question of law decided by the judge, rather than to the proximate cause stage, where the existing fact-intensive process already affords both plaintiffs and defendants a fair opportunity to present their case. As this article argued earlier, U.S. juries apply breach-foreseeability and

proximate cause-foreseeability in a manner that does not appear to systematically disadvantage plaintiffs. The problem lies at the gate, not beyond it.

5.2.2 Potential Objections and Responses

One might object that the recalibration proposed here would significantly destabilize the doctrine of duty-foreseeability. That concern is overstated. The adjustment advocated in this article applies only to a narrow class of cases – specifically, those involving personal injury caused by autonomous agents in which an AI developer is named as a defendant. It does not call for a reworking of foreseeability analysis across the entire spectrum of negligence law.

Even if the concern about doctrinal disruption were accepted, it does not follow that the law should remain static. Certain historical moments demand structural reform, and we are living through one. As Part II of this article demonstrated, the industrial revolution ushered in novel risks that prompted courts in both England and the U.S. to revise long-standing legal frameworks to meet new challenges. The technologies emerging today, including autonomous AI systems, are arguably even more complex and transformative. These systems raise legal and moral questions that are without precedent. In such contexts, the proper task is not to preserve doctrinal continuity at all costs, but to craft rules that respond meaningfully to the risks at hand. The design of legal doctrine should not be animated by an abstract commitment to doctrinal minimalism but by the nature of the problem it seeks to regulate.

One might further object that the proposed reforms would expand liability for AI developers in a way that could chill beneficial innovation by raising the costs and risks of developing and deploying autonomous agents. This concern, while legitimate, would be overstated in the present context. The proposed reforms are narrowly tailored to personal injury cases involving autonomous agents, a small subset of all AI applications. Moreover, the preservation of fact-intensive analysis at breach and proximate cause stages provides a meaningful control against over-deterrence. It would allow developers who exercise genuine care in testing, monitoring and deploying their systems to retain the opportunity to argue why they should not ultimately be held liable.

6 Conclusions

As autonomous AI agents become increasingly capable of operating without human oversight, the legal system must confront novel challenges around accountability for personal injury. This article has argued that one of the most pressing challenges lies

in the way U.S. negligence law applies the doctrine of foreseeability. When an autonomous agent causes harm and a developer is sued, traditional foreseeability analysis may appear hard to use. The features that make autonomous agents valuable, including their autonomy, efficiency and accuracy, also make it difficult to anticipate how they will behave. Yet, as this article has shown, those features are not inherent to AI agents. Rather, they are the product of abstraction choices made by developers who prioritize performance over interpretability and predictability.

These choices and the risks they create warrant corresponding doctrinal adjustments. Foreseeability should not be treated as a neutral gatekeeper in negligence law when its application serves to protect developers from the very unpredictability they engineered. Where plaintiffs suffer personal injury because of the deployment of autonomous AI agents, the law should ease the burden of foreseeability, particularly at the duty stage.

This article has proposed a three-part framework for reform. First, it recommends preserving existing foreseeability doctrine where it already operates in favor of plaintiffs. Second, it advocates for recalibrating duty-foreseeability by adopting clearer, categorical approaches that focus on general classes of harm, such as those caused by misalignment, misuse, or malfunction, rather than requiring plaintiffs to demonstrate specific pathways to injury. Third, it emphasizes the continued importance of fact-intensive analysis at the breach and proximate cause stages, where juries are best placed to assess foreseeability based on context-specific evidence.

These reforms are tailored to a narrow set of cases involving personal injury caused by Level 3 and Level 4 autonomous agents. In such cases, foreseeability should be understood as a flexible doctrine responsive to the realities of autonomous agents and grounded in the normative commitments of tort law, which include protecting bodily integrity and deterring high-risk conduct. As history has shown, courts are capable of evolving tort law to meet the demands of new technological and social circumstances. The industrial revolution prompted one such evolution, and the era of autonomous agents requires another. If the doctrine of foreseeability is to continue serving as an acceptable gate to recovery, courts must ensure that it does not become a barrier to justice.