North Carolina
Journal of Law & Technology
Volume 26, Issue 3 2025

Liability for AI Agents

Maarten Herbosch†

Artificial intelligence ("AI") is becoming integral to modern life, fueling innovation while presenting complex legal challenges. Unlike traditional software, AI operates with a degree of autonomy, producing outcomes that its developers or deployers cannot fully anticipate. Advances in underlying technology have further enhanced this autonomy, giving rise to AI agents: systems capable of interacting with their environment independently, often with minimal or no human oversight. As AI decision-making—like that of humans—is inherently imperfect, its increasing deployment inevitably results in instances of harm, prompting the critical question of whether developers and deployers should be held liable as a matter of tort law.

This question is frequently answered in the negative. Many scholars, adopting a framework of technological exceptionalism, assume AI to be uniquely disruptive. Citing the lack of transparency and unpredictability of AI models, they contend that AI challenges conventional notions of causality, rendering existing liability regimes inadequate.

This Article offers the first comprehensive normative analysis of the liability challenges posed by AI agents through a law-and-economics lens. It begins by outlining an optimal AI liability framework designed to maximize economic and societal benefits. Contrary to prevailing assumptions about AI's disruptiveness, this analysis reveals that AI largely aligns with traditional

† Assistant Professor of Law and Artificial Intelligence, Faculty of Law and Criminology, KU Leuven. Email: maarten.herbosch@kuleuven.be. I gratefully acknowledge the support of the Research Foundation Flanders (11K4421N), the Belgian American Educational Foundation, the Fulbright Commission, and Rotary D2140 for supporting this research. I also thank Oren Bar-Gill and Mason Kortz for their invaluable feedback, as well as the editors of the *NC JOLT* Executive Board. All errors remain my responsibility.

products. While AI presents some distinct challenges—particularly in its complexity, opacity, and potential for benefit externalization—these factors call for targeted refinements to existing legal frameworks rather than an entirely new paradigm.

This holistic approach underscores the resilience of traditional legal principles in tort law. While AI undoubtedly introduces novel complexities, history shows that tort law has effectively navigated similar challenges before. For example, AI's causality issues closely resemble those in medical malpractice cases, where the impact of treatment on patient recovery can be uncertain. The legal system has already addressed these issues, providing a clear precedent for extending similar solutions to AI. Likewise, while the traditional distinction between design and manufacturing defects does not map neatly onto AI, there is a compelling case for classifying inadequate AI training data as a manufacturing defect—aligning AI liability with established legal doctrine.

Taken together, this Article argues that AI agents do not necessitate a fundamental overhaul of tort law but rather call for targeted, nuanced refinements. This analysis offers essential guidance on how to effectively apply existing legal standards to this evolving technology.

Table of Contents

I. INTRODUCTION

Artificial intelligence ("AI") systems are rapidly expanding across industries, ranging from autonomous vehicles to AI-powered medical diagnostics, and driving innovation while presenting complex legal challenges. Unlike traditional software, AI systems can make decisions and execute actions in ways that are neither explicitly programmed nor entirely foreseeable by their developers or deployers.[1] This erosion of the direct nexus between human intent and system behavior complicates the application of existing tort law principles.

Moreover, as AI research advances, these systems will become increasingly autonomous, potentially culminating in AI agents capable of independently implementing their decisions across domains. As a result, scholars have increasingly questioned whether traditional legal frameworks—designed for human agency and deterministic causation—are adequate to address the complexities of AI-induced harm,[2] particularly in the case of AI agents. The inherent imperfections of AI agents—capable of causing unforeseen harm—

1. *See infra* Section II.B.
2. *See generally* Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 554–55 (2015) (discussing these challenges for robots); Margot E. Kaminski, *Technological "Disruption" of the Law's Imagined Scene: Some Lessons from* Lex Informatica, 36 BERKELEY TECH. L.J. 883, 903 (2021) (describing this phenomenon for technology on a more general level).

have drawn increased attention to liability law,[3] prompting many scholars to support a tailored accountability regime.[4]

This Article challenges the prevailing assumptions about AI's disruptive nature, such as the claim that AI's unpredictability breaches causation requirements.[5] Such assumptions are often rooted in technological exceptionalism and the corresponding belief that AI necessitates new legal frameworks.[6] Part II assesses AI agents and their properties more thoroughly while considering the role technological exceptionalism plays in their perception. Rather than adopting that perception blindly, Part III offers the first comprehensive normative assessment of AI agent liability through a law-and-economics lens. This assessment suggests that AI's distinct challenges arise primarily in

---

3. *See, e.g.*, Marguerite E. Gerstner, *Liability Issues with Artificial Intelligence Software*, 33 SANTA CLARA L. REV. 239, 239 (1993); David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117, 129–50 (2014); Yavar Bathaee, *Artificial Intelligence Opinion Liability*, 35 BERKELEY TECH. L.J. 113, 154–70 (2020); Anat Lior, *AI Strict Liability Vis-À-Vis AI Monopolization*, 22 COLUM. SCI. & TECH. L. REV. 90, 94–106 (2020); Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. 1315, 1318–76 (2020); Mihailis E. Diamantis, *Employed Algorithms: A Labor Model of Corporate Liability for AI*, 72 DUKE L.J. 797, 812–16 (2023); Kathryn Bosman Cote, *Outsmarting Smart Devices: Preparing for AI Liability Risks and Regulations*, 25 SAN DIEGO INT'L L.J. 101, 119–26 (2024).

4. *See, e.g.*, 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, 90–114 (2019) (arguing in favor of the development of a specific regime for some AI systems); Alicia Lai, *Artificial Intelligence, LLC: Corporate Personhood as Tort Reform*, 2021 MICH. ST. L. REV. 597, 631–53 (2021) (proposing personhood for AI systems); Sahara Shrestha, *Nature, Nurture, or Neither?: Liability for Automated and Autonomous Artificial Intelligence Torts Based on Human Design and Influences*, 29 GEO. MASON L. REV. 375, 401–10 (2021) (proposing a new balancing test to determine liability); Renee Henson, *"I Am Become Death, the Destroyer of Worlds": Applying Strict Liability to Artificial Intelligence as an Abnormally Dangerous Activity*, 96 TEMP. L. REV. 349, 362–90 (2024) (proposing an extension of the "abnormally dangerous activities" test under products liability).

Alternatively, some authors argue that AI itself requires change to be adequately captured by the legal framework. *See* Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 COLUM. L. REV. 1829, 1832–50 (2019) (stressing the need for explainability).

5. *See infra* Section IV.D.

6. *See infra* Section II.C.

highly complex systems and those with increased benefit externalization, rather than across the board. Part IV subsequently examines the existing legal framework for AI, highlighting persistent issues such as causality requirements and the distinction between manufacturing and design defects. This Part also demonstrates that similar challenges exist in traditional contexts, underscoring the ability of tort law to address such issues without requiring a fundamental overhaul. By integrating these insights, Part V proposes targeted modifications to enhance tort law's ability to address AI-induced harm while encouraging the responsible deployment of AI agents for societal benefit.

Rather than centering on specific AI applications, this Article takes a broad approach, enabling a comprehensive analysis of AI liability challenges across various sectors. While some literature focuses on areas like autonomous vehicles or medical AI—yielding precise but limited insights[7]—such a narrow focus risks overlooking broader implications. It also restricts the discussion mainly to physical harm,[8] despite such cases representing only a small fraction of AI applications.[9]

A holistic legal framework must account for varying degrees of control among affected parties. Because victims of AI-induced harm may or may not be system deployers, they can have differing levels of influence over AI use. Examining both extremes—such as a driver choosing to utilize an autonomous vehicle versus a patient being unknowingly subjected to AI in a hospital—underscores the diverse liability considerations and varying levels of control[10] that a comprehensive framework must address.

---

7. These areas also present challenges for the liability framework in the absence of AI. *See, e.g.*, Michael L. Rustad & Thomas H. Koenig, *Cybertorts and Legal Lag: An Empirical Analysis*, 13 S. CAL. INTERDISC. L.J. 77, 77 (2003) (describing how the body of automobile law has quickly expanded since the start of the twentieth century).

8. Selbst, *supra* note 3, at 1318.

9. *Id.* at 1318–19 (discussing autonomous robots used in medical practice and autonomous vehicles).

10. *See, e.g.*, Madeline Roe, *Who's Driving That Car?: An Analysis of Regulatory and Potential Liability Frameworks for Driverless Cars*, 60 B.C. L. REV. 317, 337 (2019)

footnote continued on next page

II. AI AGENTS

A. Artificial Intelligence & AI Agents

AI, like many technological concepts, is challenging to define. Over the years, various definitions have emerged. Initially, researchers described AI as the simulation of human intelligence by computer systems.[11] However, given the capabilities of both contemporary AI systems and "traditional" non-AI computer systems, this definition has grown increasingly inadequate. Some alternative definitions focus on specific computer techniques characterizing AI,[12] but these lack technological neutrality and require constant updating. From a legal perspective, defining AI by its increasing autonomy offers more practicality.[13] Thus, AI systems can be defined as computer systems designed to operate with varying levels of autonomy and potential adaptiveness. Here, autonomy is understood in a non-philosophical sense,[14] meaning that the behavior

(discussing surgical robots and autonomous vehicles); *cf. Miriam C. Buiten, Product Liability for Defective AI*, 57 EUR. J.L. & ECON. 239, 244 (2024).

11. J. McCarthy et al., A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence 2 (Aug. 31, 1955) (unpublished research proposal) (on file with Stanford University); *cf.* Maura R. Grossman & Gordon V. Cormack, *The Grossman-Cormack Glossary of Technology-Assisted Review*, 7 FED. CTS. L. REV. 85, 87 (2014); Shlomit Yanisky-Ravid, *Generating Rembrandt: Artificial Intelligence, Copyright, and Accountability in the 3A Era—The Human-Like Authors Are Already Here—a New Model*, 2017 MICH. ST. L. REV. 659, 673 (2017); Mircea-Constantin Şcheau et al., *Artificial Intelligence/Machine Learning Challenges and Evolution*, 7 INT'L J. INFO. SEC. & CYBERCRIME 11, 12 (2018); Ronald Yu & Gabriele Spina Alì, *What's Inside the Black Box? AI Challenges for Lawyers and Researchers*, 19 LEGAL INFO. MGMT. 2, 2 (2019).

12. See in particular Article 3 (1) of the original (not adopted) *Commission Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, tit. I, art. 3(1), COM (2021) 206 final (Apr. 21, 2021), as well as *id.* at Annex I.

13. *Cf.* Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), art. 3 (1), 2024 O.J. (L 1689) 1 [hereinafter AI Act].

14. For an example of a more philosophical meaning of "autonomy," see S. I. Benn, *Freedom, Autonomy and the Concept of a Person*, 76 PROCS. ARISTOTELIAN SOC'Y 109, 124 (1976); John Christman, *Liberalism, Autonomy, and Self-Transformation*, 27 SOC. THEORY & PRAC. 185, 187–190 (2001).

of modern AI systems is not entirely predetermined by their developers or deployers.[15]

AI systems can autonomously generate outputs based on inputs, fundamentally differing from traditional computer systems. In traditional algorithms, programmers specify exact responses for given inputs (e.g., "if *X*, then *Y*").[16] This is similar to a vending machine's operation: If a coin is inserted and a button is pressed, a drink is dispensed. AI systems, particularly those using machine learning, illustrate a significant contrast.[17] For instance, an AI spam filter trained on examples of spam and regular emails develops its own classification rules for new emails, potentially eliminating the need for developers to program these rules explicitly.[18] This autonomy sets AI systems apart from traditional computer systems.

The concept of autonomy is central to the idea of AI agents, which broadly refers to AI systems deployed with minimal or no human oversight. AI agents operate independently, interacting with their environment alongside varying degrees of human intervention. This Article adopts a broad definition of AI agents, including those with some oversight. An effective liability regime must be able to address both fully autonomous and partially supervised systems.

The scope of this Article is restricted to existing and foreseeable AI systems and agents. It is thus useful to stress that those constitute what is generally known as "weak" AI, capable of performing specific, targeted tasks but lacking the general intelligence and complete

15. While 15 U.S.C. § 9401(3) does not explicitly refer to autonomy, it does implicitly appear to enshrine it by referring to "automated" analysis.

16. *Cf.* Pompeii Ests., Inc. v. Consol. Edison Co., 397 N.Y.S.2d 577, 579 (N.Y. Civ. Ct. 1977) ("Computers can only issue mandatory instructions."); Gerstner, *supra* note 3, at 239.

17. *See, e.g.*, Weston Kowert, *The Foreseeability of Human–Artificial Intelligence Interactions*, 96 TEX. L. REV. 181, 183 (2017) (discussing the use of AI for autonomous vehicles).

18. Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 GA. ST. U. L. REV. 1305, 1314 (2019). *See generally* Emmanuel Gbenga Dada et al., *Machine Learning for Email Spam Filtering: Review, Approaches, and Open Research Problems*, 5 HELIYON 1, 2–4 (2019) (providing a technical description of AI-powered spam filters).

autonomy of humans.[19] Conversely, "strong" AI[20]—also known as artificial general intelligence ("AGI")[21]—would be capable of replicating human-like intelligence.[22] However, strong AI remains a hypothetical concept, as it does not yet exist.[23]

Before examining the tort law regime applicable to AI agents, both normatively and descriptively, it is crucial to first highlight certain fundamental characteristics of these systems. By doing so, one avoids oversimplifying the analysis and allows for a holistic and nuanced understanding of the technology. Additionally, these characteristics are essential for assessing the challenges AI agents present under a tort law framework.

19. TED GOERTZEL, *Path to More General Artificial Intelligence*, *in* RISKS OF ARTIFICIAL INTELLIGENCE 69, 70 (Vincent C. Müller ed., 2015).

20. *See, e.g.*, George S.K., *Can Artificial Intelligence Machines Be Patented or Sued?*, 6 CT. UNCOURT 41, 41 (2019).

21. *See, e.g.*, Jason Chung & Amanda Zink, *Hey Watson, Can I Sue You for Malpractice? Examining the Liability of Artificial Intelligence in Medicine*, 11 ASIA PAC. J. HEALTH L. & ETHICS 51, 53 (2017); Selbst, *supra* note 3, at 1344.

22. Roni A. Elias, *Facing the Brave New World of Killer Robots: Adapting the Development of Autonomous Weapon Systems into the Framework of the International Law of War*, 21 TRINITY L. REV. 70, 87 (2016); Brian L. Frye, *The Lion, the Bat & the Thermostat: Metaphors of Consciousness*, 5 SAVANNAH L. REV. 13, 19 (2018); Stephen E. Henderson, *Should Robots Prosecute and Defend?*, 72 OKLA. L. REV. 1, 3 (2019).

23. Elias, *supra* note 22, at 72; JOHN BUYERS, ARTIFICIAL INTELLIGENCE: THE PRACTICAL LEGAL ISSUES 6 (2018); Frye, *supra* note 22, at 19; John Frank Weaver, *Everything Is Not Terminator: America's First AI Legislation*, 1 J. ROBOTICS A.I. & L. 201, 202 (2018); Henderson, *supra* note 22, at 7–8; John Linarelli, *Artificial General Intelligence and Contract*, 24 UNIF. L. REV. 330, 331 (2019); Wim Naudé & Nicola Dimitri, *The Race for an Artificial General Intelligence: Implications for Public Policy*, 35 AI & SOC'Y 367, 368 (2020); Patric M. Reinbold, *Taking Artificial Intelligence Beyond the Turing Test*, 2020 WIS. L. REV. 873, 874 (2020); Robin Feldman & Kara Stein, *AI Governance in the Financial Industry*, 27 STAN. J.L. BUS. & FIN. 94, 102 (2022). Some authors doubt whether this will ever be possible. *See, e.g.*, John O. McGinnis, *Accelerating AI*, *in* RESEARCH HANDBOOK ON THE LAW OF ARTIFICIAL INTELLIGENCE 369 (Woodrow Barfield & Ugo Pagallo eds., 2018).

B. AI System Characteristics

1. Performance

The autonomy of AI systems and agents is a key factor driving their popularity.[24] AI systems can often outperform humans in various domains[25] as they are not constrained by conceptual links between inputs and outputs.[26] Instead, they identify statistical correlations to optimize their output.[27]

AI systems and agents also benefit from many traditional advantages of computer systems. Automation often results in faster and cheaper task execution than human performance would allow.[28]

---

24. For a discussion on the desirability of such autonomy, see Curtis E.A. Karnow, *Liability for Distributed Artificial Intelligences*, 11 BERKELEY TECH. L.J. 147, 154 (1996).

25. *Cf.* Ryan Calo, *Singularity: AI and the Law*, 41 SEATTLE U. L. REV. 1123, 1124 (2018); Michael Hatfield, *Professionally Responsible Artificial Intelligence*, 51 ARIZ. ST. L.J. 1057, 1060 (2019); Wojciech Samek & Klaus-Robert Müller, *Towards Explainable Artificial Intelligence*, *in* EXPLAINABLE AI: INTERPRETING, EXPLAINING AND VISUALIZING DEEP LEARNING 5, 5–6 (Wojciech Samek et al. eds., 2019). This is beautifully exemplified by AlphaGo Zero, an AI system developed to master the board game "Go." It was long assumed this game was too complicated for computer systems to master, but AlphaGo Zero can now beat the best human players consistently. *See* Alan Levinovitz, *The Mystery of Go, the Ancient Game That Computers Still Can't Win*, WIRED (May 12, 2014, 6:30 AM), https://www.wired.com/2014/05/the-world-of-computer-go/ [https://perma.cc/PUV7-RP77]; David Silver et al., *Mastering the Game of Go Without Human Knowledge*, 550 NATURE 354, 354 (2017).

26. As a result, AI systems do not "think" like humans. *See* Daniel Martin Katz, *Quantitative Legal Prediction—Or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 EMORY L.J. 909, 918 (2013); Iria Giuffrida et al., *A Legal Perspective on the Trials and Tribulations of AI: How Artificial Intelligence, the Internet of Things, Smart Contracts, and Other Technologies Will Affect the Law*, 68 CASE W. RSRV. L. REV. 747, 755 (2018); Surden, *supra* note 18, at 1315.

27. *Cf.* KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS: NEW TOOLS FOR LAW PRACTICE IN THE DIGITAL AGE 234 (2017); Giuffrida et al., *supra* note 26, at 766. For a discussion on supervised learning, see GOPINATH REBALA ET AL., AN INTRODUCTION TO MACHINE LEARNING 20 (2019), and Surden, *supra* note 18, at 1314.

28. *See also* Cass v. 1410088 Ontario Inc., 2018 ONSC 6959, para. 34 (Can. Ont. Sup. Ct. J.). *Cf.* Harry Surden, *Computable Contracts*, 46 U.C. DAVIS L. REV. 629, 638 (2012).

However, this advantage is slightly nuanced for AI systems, as they may require significant investments for optimization, including high-quality training data, advanced hardware, and a well-performing algorithm.[29]

Despite their strengths, however, AI systems are inherently imperfect,[30] and errors are unavoidable if a system is used frequently enough.[31] As with traditional software, these errors may stem from bugs caused by developer oversight.[32] An intriguing example is "reward function hacking,"[33] where an AI system achieves its objectives in unexpected ways—such as a robot vacuum ejecting dust solely to collect it again.[34]

Additionally, AI systems often operate in probabilistic environments where outputs are statistically determined.[35] This can

29. For an interesting illustration of what these costs may look like for a specific AI tool, see Matt Tanner, *The Cost of Building AI: Understanding AI Cost Analysis*, MOESIF BLOG (July 25, 2024), https://www.moesif.com/blog/technical/api-development/The-Cost-of-Building-AI-Understanding-AI-Cost-Analysis/ [https://perma.cc/G7WZ-ZACG].

30. Nathan A. Greenblatt, *Self-Driving Cars and the Law*, 53 IEEE SPECTRUM 46, 48 (2016). *Cf.* Azim Shariff et al., *Psychological Roadblocks to the Adoption of Self-Driving Vehicles*, 1 NATURE HUM. BEHAV. 694, 695 (2017); Bryan H. Choi, *Crashworthy Code*, 94 WASH. L. REV. 39, 86 (2019) (discussing computer systems more generally). This statement also applies more generally to products, sometimes discussed as "unavoidable harms." *See, e.g.*, Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 U. CHI. L. REV. 1311, 1327–28 (2019).

31. Greenblatt, *supra* note 30, at 48. *Cf.* Chagal-Feferkorn, *supra* note 4, at 64, 84; Selbst, *supra* note 3, at 1318.

32. Gerstner, *supra* note 3, at 246. *Cf.* Selbst, *supra* note 3, at 1325 n.33 (referring to the Halting problem); Deven R. Desai & Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH 1, 31–32 (2017).

33. *See generally* Yinlong Yuan et al., *A Novel Multi-Step Reinforcement Learning Method for Solving Reward Hacking*, 49 APPLIED INTEL. 2874, 2874 (2019) (discussing reward hacking).

34. STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 37 (3d ed. 2010); Dylan Hadfield-Menell et al., *Inverse Reward Design*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 1, 1 (Isabelle Guyon et al. eds., 2017).

35. *See* Chagal-Feferkorn, *supra* note 4, at 84.

lead to some conceptually surprising errors.[36] For example, an AI system may fail to recognize an image because a few pixels were altered, even if the change is imperceptible to the human eye.[37] Other sources of error can include poor system training, which can result from the use of inferior training data (e.g., high-quality data that contains embedded societal biases) or simply an insufficient quantity of training data.[38] Errors may also arise when an AI system encounters situations outside the scope of its training data.[39]

Developers and deployers can mitigate errors and improve system accuracy through diligent programming and adequate training. However, perfection remains unattainable.[40] Efforts to achieve marginal performance gains require drastically increasing investments in data and time,[41] eventually making the costs disproportionate to the benefits.[42] This tradeoff is a critical factor for the normative assessment in Part III below, as it underscores the impracticality of requiring AI developers to create flawless systems.

---

36. *E.g.*, Steve Lohr, *Facial Rexognition Is Accurate, if You're a White Guy*, N.Y. TIMES (Feb. 9, 2018), https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html [https://perma.cc/H5JP-XY5P].

37. Anh Nguyen et al., *Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images*, PROCEEDINGS OF THE IEEE CONFERENCE ON COMPUTER VISION & PATTERN RECOGNITION 427 (2015). *Cf.* Feldman & Stein, *supra* note 23, at 103 (discussing a change that is noticeable but still yields disproportionately incorrect results).

38. *See generally* Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 105–07 (2014) (discussing limitations of legal predictive models). *See also e.g.*, Brian Jalaian et al., *Uncertain Context: Uncertainty Quantification in Machine Learning*, AI MAG., Winter 2019, at 45 (noise in data).

39. Roe, *supra* note 10, at 338; *see also* Frank Griffin, *Artificial Intelligence and Liability in Health Care*, 31 HEALTH MATRIX 65, 79 (2021).

40. *See* sources cited *supra* note 30. This challenge is also present for more traditional computer systems: the system is never perfect, but more testing can help spot increasingly fewer bugs. *See* Choi, *supra* note 30, at 86–87.

41. *See generally* sources cited *supra* note 30. Neil C. Thompson et al., *Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable*, 58 IEEE SPECTRUM 50, 55 (2021). Such diminishing returns are a necessary consequence of the inherent imperfection of AI systems. *See* Andrew Majot & Yampolskiy Roman, *Diminishing Returns and Recursive Self-Improving Artificial Intelligence*, *in* THE TECHNOLOGICAL SINGULARITY 141, 148 (Vincent Callaghan et al. eds., 2017).

42. *See generally* sources cited *supra* notes 30, 41.

2. Foreseeability and Explainability

As with the performance of AI systems, the foreseeability and intelligibility of AI outputs are also legally significant. First, as discussed earlier, AI outputs are often unforeseeable.[43] It is largely unpredictable how an AI system or agent might respond to certain inputs,[44] particularly for systems that continue to learn during operation.[45]

Second, the mathematical, rather than conceptual,[46] nature of AI decision-making, often involving intricate and large-scale computations, can make outputs difficult to understand.[47] This challenge is especially pronounced in complex systems such as deep learning algorithms, which are often referred to as "black boxes."[48]

Importantly, explainability is relative, depending on the audience attempting to understand the system and the depth of understanding

43. *Cf.* Lemley & Casey, *supra* note 30, at 1334 (discussing how this also applies to more traditional, non-AI systems, and how it is a more prominent and essential trait here).

44. *See* Kowert, *supra* note 17, at 183; Yu & Ali, *supra* note 11, at 5; Selbst, *supra* note 3, at 1332; William D. Smart et al., *An Education Theory of Fault for Autonomous Systems*, 2 NOTRE DAME J. EMERGING TECH. 33, 35 (2021); Matthew Oliver, *Contracting by Artificial Intelligence: Open Offers, Unilateral Mistakes, and Why Algorithms Are Not Agents*, 2 ANU J.L. & TECH. 45, 50 (2021).

45. *See* Selbst, *supra* note 3, at 1332. *See also* Kowert, *supra* note 17, at 184.

46. W. Nicholson Price & Arti K. Rai, *Clearing Opacity through Machine Learning*, 106 IOWA L. REV. 775, 778–79 (2021).

47. Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 889, 897 (2017); Hatfield, *supra* note 25, at 1118 n.278; Samek & Müller, *supra* note 25, at 6; Yu & Ali, *supra* note 11, at 5; Adrien Bibal et al., *Legal Requirements on Explainability in Machine Learning*, A.I. & L. 1, 10 (2020); Alicia Solow-Niederman, *Administering Artificial Intelligence*, 93 S. CAL. L. REV. 633, 657 (2020); Price & Rai, *supra* note 46, at 779; Amy L. Stein, *Assuming the Risks of Artificial Intelligence*, 102 B.U. L. REV. 979, 1005 (2022). *Cf.* Charlotte A. Tschider, *Medical Device Artificial Intelligence: The New Tort Frontier*, 46 BYU L. REV. 1551, 1560 (2020). This even applies to the system developer, see *id.* at 1560–61. *Cf.* Selbst, *supra* note 3, at 1334.

48. Bathaee, *supra* note 47, at 897; Hatfield, *supra* note 25, at 1118 n.278; Samek & Müller, *supra* note 47, at 6; Yu & Ali, *supra* note 11, at 5; Solow-Niederman, *supra* note 47, at 657.

they require.[49] Explainability arguably exists on a spectrum; a system or agent is not simply "explainable" or "not explainable." Instead, it can be more or less explainable based on how conceptually accessible its "motives" are and the effort and expertise required to interpret them.[50]

A lack of adequate explainability or interpretability limits the supervision developers or deployers can exercise. When the AI agent's "reasoning" process cannot be verified, oversight is reduced to checking outputs using alternative, human-conceptual methods if they are available.[51] However, this does not ensure that the agent arrived at a given output "for the right reasons."[52] This lack of transparency becomes particularly problematic if there are no human-conceptual

---

49. The latter is captured by the terminological confusion between "interpretability" and "explainability". Both notions are sometimes used interchangeably. Bennoush Abdollahi & Olfa Nasraoui, *Transparency in Fair Machine Learning: the Case of Explainable Recommender Systems*, *in* HUMAN AND MACHINE LEARNING: VISIBLE, EXPLAINABLE, TRUSTWORTHY AND TRANSPARENT 21, 24 (Jianlong Zhou et al. eds., 2018); PATRICK HALL & NAVDEEP GILL, AN INTRODUCTION TO MACHINE LEARNING INTERPRETABILITY 2 (2019); Mengnan Du et al., *Techniques for Interpretable Machine Learning*, 63 COMM. ACM 68, 69 (2020); Gianfagna & Di Cecco, *supra* note 58, at 25. They are sometimes used to denote different levels to describe the degree to which the conceptual process is clear, with "explainability" indicating a higher threshold than "interpretability." Hall & Gill, *supra*, at 2; LEONIDA GIANFAGNA & ANTONIO DI CECCO, EXPLAINABLE AI WITH PYTHON 12 (2021).

50. *Cf.* Hadji Misheva, Branka et al., Audi*ence-Dependent Explanations for AI-Based Risk Management Tools: A Survey*, 4 FRONT. ARTIFICIAL INTELLIGENCE 794996, 6 (2021).

51. *Cf. id.* This may be particularly problematic in instances where the deployer does not have the necessary expertise. A doctor (as in Example 1, *infra* Part III) could still rely on their own training—but that does not similarly apply to someone with any relevant expertise (such as the client in Example 3, *infra* Part III).

52. Mateusz Szczepański et al., *The Methods and Approaches of Explainable Artificial Intelligence*, *in* COMPUTATIONAL SCIENCE: ICCS 3, 4–5 (Maciej Paszynski et al. eds., 2021). This is beautifully illustrated by the "Clever Hans Effect," describing a horse (Hans) that was attributed significant intelligence as it seemed able to perform calculations and spell words. A subsequent analysis revealed that the horse was not actually performing those cognitive tasks, but was rather picking up on the body language of the person asking the question. *See id.*

methods available to verify a system's procedure in reaching a given output.[53]

As a result, significant attention is devoted to "explainable AI" in research.[54] This can be approached in two main ways. First, developers may opt for AI models that are inherently more explainable.[55] However, given the widespread use of complex models, such as large language models and neural networks, a second approach—*post-hoc* explainability—has gained prominence.[56] This method adds a layer of interpretability to models that would otherwise be "black boxes,"[57] employing techniques such as counterfactual explanations.[58] These are, however, not always sufficient to fully understand the system's decision-making process.[59]

C. Legal Analysis

1. Qualification

AI systems or agents are not subject to a distinct legal regime; they are treated as tools under current law, much like traditional computer systems. However, some scholars advocate granting AI systems and

53. This might make it, for example, impossible to rule out that the system obtained the relevant output based on undesirable biased assumptions. Du et al., *supra* note 49, at 69.

54. Amina Adadi & Mohammed Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, 6 IEEE ACCESS 52138, 52139 (2018); Deeks, *supra* note 4, at 1833–34; Stein, *supra* note 47, at 1005–06. *See also* Paulo Henrique Padovan et al., *Black Is the New Orange: How to Determine AI Liability*, 31 A.I. & L. 133, 151–58 (2023).

55. *Cf.* Du et al., *supra* note 49, at 75; Arun Rai, *Explainable AI: From Black Box to Glass Box*, 48 J. ACAD. MKTG. SCI. 137, 138 (2020) (implicitly); UDAY KAMATH & JOHN LIU, EXPLAINABLE ARTIFICIAL INTELLIGENCE: AN INTRODUCTION TO INTERPRETABLE MACHINE LEARNING 16 (2022).

56. Rai, *supra* note 55, at 138. Those more complex models are inherently less explainable, thus excluding inherent explainability—leaving only post-hoc options. *See id.*

57. *Id.*; KAMATH & LIU, *supra* note 55, at 17.

58. Du et al., *supra* note 49, at 71; Rai, *supra* note 55, at 138; GIANFAGNA & DI CECCO, *supra* note 49, at 20; KAMATH & LIU, *supra* note 55, at 17.

59. Du et al., *supra* note 49, at 75 (describing how post-hoc explainability approximates that decision-making process but may fail to do so adequately accurately—particularly for more complex models).

agents limited legal personhood based on their unique properties,[60] thereby holding the agents themselves accountable for damages the systems cause.[61] While seemingly straightforward, such proposals introduce significant complexities: the need to define AI legal capacity, establish an estate for such agents, and ensure human representation.[62] The last of these raises profound conceptual challenges, making this approach far from simple.

Modern AI agents demonstrate impressive autonomy, which is the primary argument for granting them some form of legal personality.[63] However, their autonomy remains inherently limited. Without strong AI, these systems and agents cannot function independently in many contexts, necessitating human representation.[64] This limitation weakens the appeal of recognizing AI agents as "e-persons."

2. Technological Exceptionalism

In many ways, the argument for recognizing AI agents as legal entities underscores the influence of technological exceptionalism in this field. Exceptionalism often serves as a foundation to understand and guide policy on technological development,[65] premised on the idea that certain technologies necessitate profound changes to laws or institutions in order to uphold societal justice.[66]

---

60. Vladeck, *supra* note 3, at 121 (discussing strong AI); Anat Lior, *AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy*, 46 MITCHELL HAMLINE L. REV. 1043, 1065–71 (2020); *see also* Giuffrida et al., *supra* note 26, at 765 (descriptively).

61. *Cf.* Chung & Zink, *supra* note 21, at 69–70. Similarly, this Article would argue that a discussion of vicarious liability regimes as it applies to AI agents, *see, e.g.*, Sam N. Lehman-Wilzig, *Frankenstein Unbound: Towards a Legal Definition of Artificial Intelligence*, 13 FUTURES 442, 451–52 (1981), is equally misplaced.

62. *See* Malcolm Bain & Brian Subirana, *Legalising Autonomous Shopping Agent Processes*, 19 COMPUT. L. & SEC. REV. 375, 376 (2003); Tina Balke & Torsten Eymann, *The Conclusion of Contracts by Software Agents in the Eyes of the Law*, *in* 7TH INTERNATIONAL CONFERENCE ON AUTONOMOUS AGENTS AND MULTI AGENT SYSTEMS 771, 773 (Lin Padgham et al. eds., 2008).

63. *See, e.g.*, Bain & Subirana, *supra* note 62, at 377; Vladeck, *supra* note 3, at 124.

64. *See* sources cited *supra* note 62.

65. *See* Meg Leta Jones, *Does Technology Drive Law? The Dilemma of Technological Exceptionalism in Cyberlaw*, 2018 U. ILL. J.L. TECH. & POL'Y 249, 284 (2018).

66. *See also* Calo, *supra* note 2, at 103; Jones, *supra* note 65, at 253 (more implicitly).

However, exceptionalism offers limited guidance on addressing the normative balance required by AI and similar technologies. Instead, it highlights areas requiring attention, emphasizing the disruptive nature of sufficiently "new" or challenging technologies.[67] These technologies are often portrayed as evading existing legal frameworks or destabilizing their established balances.[68]

Similarly, it is often argued that new technology cannot be regulated proactively because policymakers may lack a full understanding of its properties or potential.[69] This creates a dichotomy: technologies that are not deemed sufficiently new or disruptive are considered *manageable* within existing legal frameworks, while those perceived as sufficiently novel and disruptive are said to *necessitate* new legislation.[70]

Exceptionalism views the relationship between novel technologies and the law as a linear progression: a legal framework exists, a technological innovation emerges, and this disrupts the framework such that it requires legal adaptation.[71] In this perspective, technology

67. *See, e.g.*, Joshua Schoonmaker, *Proactive Privacy for a Driverless Age*, 25 INFO. & COMMC'NS TECH L. 96, 97 (2016).

On AI as "disruptive," *see* Mariano-Florentino Cuéllar, *Cyberdelegation and the Administrative State*, *in* ADMINISTRATIVE LAW FROM THE INSIDE OUT: ESSAYS ON THEMES IN THE WORK OF JERRY L. MASHAW 134, 237 (Nicolas R. Parillo ed., 2017); Alessandro Miasato & Fabiana Reis Silva, *Artificial Intelligence as an Instrument of Discrimination in Workforce Recruitment*, 8 ACTA UNIV. SAPIENTIAE: LEGAL STUD. 191, 193–94 (2019); Horst Eidenmuller & Faidon Varesis, *What Is an Arbitration? Artificial Intelligence and the Vanishing Human Arbitrator*, 17 N.Y.U. J.L. & BUS. 49, 51 (2020); *see also* Kaminski, *supra* note 2, at 892–95 (discussing legal disruption more generally).

68. *See, e.g.*, Schoonmaker, *supra* note 67, at 97 ("This sort of technological disruption is shaking a number of legal realms as they struggle to keep up with the relentless pace of innovation.").

69. Jones, *supra* note 65, at 250.

70. *See id.*

71. *Id.* at 250–51 (2018). *See, e.g.*, Calo, *supra* note 2, at 556–57. *Cf.* Arthur Cockfield & Jason Pridmore, *A Synthetic Theory of Law and Technology*, 8 MINN. J.L. SCI. & TECH. 475, 476, 489 (2007).

is seen as the force that "drives" the law,[72] closely tied to the concept of the "pacing problem."[73]

Technological exceptionalism gained prominence with the advent of the internet, as evidenced by the strong pushback against Frank Easterbrook's "law of the horse"[74] analogy.[75] Consequently, it has become the prevailing legal framework for addressing novel technology.[76]

Despite its prominence, exceptionalism has faced criticism on multiple fronts.[77] It has often proven inadequate in addressing past technological innovations.[78] Moreover, it tends to undervalue human agency,[79] overlooking the ability of individuals to comprehend and manage the limitations of technological systems.[80] Additionally, exceptionalism can distort perceptions of technology by overstating challenges. For instance, some legal literature on AI exaggerates the

---

72. For critical reflections, see Jones, *supra* note 65, at 249; Kaminski, *supra* note 2, at 895.

73. *Cf.* Joel R. Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules through Technology*, 76 TEX. L. REV. 553, 586 (1998) ("[T]oday's regulations may easily pertain to yesterday's technologies."); Gary E. Marchant, THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND THE LEGAL-ETHICAL OVERSIGHT 19, 22–23 (Gary E. Marchant et al. eds., 2011); Jones, *supra* note 65, at 251.

74. Frank H. Easterbrook, *Cyberspace and the Law of the Horse*, 1996 U. CHI. LEGAL F. 207 (1996).

75. *See, e.g.*, Lawrence Lessig, *The Law of the Horse: What Cyber Law Might Teach*, 113 HARV. L. REV. 501, 546 (1999); *cf.* Kaminski, *supra* note 2, at 883 (more implicitly).

76. Jones, *supra* note 65, at 251–54 (discussing technological determinism more generally); *cf.* Gaia Bernstein, *Toward a General Theory on Law and Technology: Introduction*, 8 MINN. J.L. SCI. & TECH. 441, 441–42 (2007).

77. *See* Jones, *supra* note 65, at 249; Kaminski, *supra* note 2, at 895.

78. *Cf.* SHEILA JASANOFF, SCIENCE AT THE BAR 22 (1995) (implicitly); Jones, *supra* note 65, at 260, 284.

79. In this sense, it corresponds (to some degree) to substantive theories of technology that emphasize how individuals may be impacted by technology. *Cf.* Cockfield & Pridmore, *supra* note 71, at 475–76 (2007).

80. *Id.* at 480.

unpredictability or lack of explainability of AI, essentially attributing traits of AGI.[81] These issues will be revisited later in this Article.[82]

Criticisms of technological exceptionalism, however, do not diminish its value. Instead, these criticisms acknowledge that it plays a crucial role in identifying challenges posed by novel technologies and highlighting areas requiring significant or immediate legal reform.[83] Nevertheless, the law is often more "future-proof"[84] than exceptionalism might imply.[85] While the pacing problem and exceptionalism may fully apply in some aspects of technology policy,[86] they should neither be the exclusive nor dominant approach,[87] as instances of fundamental disruption remain rare.[88]

Incorporating constructivism offers a more balanced and, in turn, more compelling perspective. The law frequently incentivizes and guides technological development and innovation.[89] Moreover, addressing innovation within traditional legal frameworks enhances legal certainty and maintains coherence. With this in mind, the application of existing tort law to AI agents must be assessed by examining whether and where the legal framework falls short, thereby enabling a thoughtful consideration of the scope and impact of potential legal reforms.

---

81. *See, e.g.*, Bathaee, *supra* note 47, at 924 (claiming a causation challenge exists because AI's decisions may be unforeseeable by its creators or users). Such claims may apply for AGI. For existing systems, however, the type of output is generally foreseeable, thus not obstructing an application of causation, *infra* Section IV.C.1.c.

82. *Infra* Section IV.D.

83. *Cf.* LAWRENCE LESSIG, CODE: AND OTHER LAWS OF CYBERSPACE VERSION 2.0 25 (2006); Calo, *supra* note 2, at 552; Kaminski, *supra* note 2, at 892.

84. *See, e.g.*, Kaminski, *supra* note 2, at 891.

85. *See* Jones, *supra* note 65, at 278; Kaminski, *supra* note 2, at 891.

86. *See* Kaminski, *supra* note 2, at 892.

87. *Cf.* Tim Wu, *Is Internet Exceptionalism Dead?*, *in* THE NEXT DIGITAL DECADE: ESSAYS ON THE FUTURE OF THE INTERNET 179–80 (Berin Srzoka & Adam Marcus eds., 2011) (discussing the internet); Jones, *supra* note 65, at 251; *id.* at 280.

88. *Cf.* Calo, *supra* note 2, at 558.

89. Kaminski, *supra* note 2, at 892.

III. NORMATIVE AI LIABILITY

Effectively assessing the application of tort law to AI agents requires a robust normative framework. To this end, this Article utilizes the positive economic theory of tort law, which analyzes tort law by emphasizing the optimal incentives a liability regime should provide from an economic perspective.[90] The goal is to identify these incentives to ensure tort law effectively encourages optimal behavior that maximizes societal benefits.[91]

An economic approach can help determine the appropriate level of care various parties in a potential injury case should maintain, aiming to minimize societal costs associated with both the burden of care and resulting injuries.[92] This approach is particularly justified, as numerous scholars have employed economic analyses in evaluating tort law's application to AI agents.[93] Although these analyses may appear theoretical, they offer valuable insights into the priorities a liability regime should emphasize, such as incentivizing developers, manufacturers, and operators to adopt safer practices.[94]

By imposing liability, tort law serves as a mechanism to align individual behaviors with socially optimal outcomes, reinforcing its relevance in addressing the challenges posed by AI. Some considerations are critical in this normative assessment. First, while AI may outperform humans in some respects, it also carries the risk of making errors. Its desirability depends on context. For example, it is

---

90. *See, e.g.*, Oren Bar-Gill & Ariel Porat, *Harm–Benefit Interactions*, 16 AM. L. & ECON. REV. 86, 87 (2013).

91. *See, e.g.*, *id.*

92. *Cf.* William M. Landes & Richard A. Posner, *The Positive Economic Theory of Tort Law*, 15 GA. L. REV. 851, 868 (1980); STEVEN SHAVELL, FOUNDATIONS OF ECONOMIC ANALYSIS OF LAW 182 (2004).

93. *See, e.g.*, Kevin Funkhouser, *Paving the Road Ahead: Autonomous Vehicles, Products Liability, and the Need for a New Approach*, 2013 UTAH L. REV. 437, 454 (2013); Vladeck, *supra* note 3, at 148; Helen Smith & Kit Fotheringham, *Artificial Intelligence in Clinical Decision-Making: Rethinking Liability*, 20 MED. L. INT'L 131, 148 (2020); Lai, *supra* note 4, at 618; Vincent R. Johnson, *Artificial Intelligence and Legal Malpractice Liability*, 14 ST. MARY'S J. LEGAL MALPRACTICE & ETHICS 55, 62 (2024) (implicitly).

94. *Cf.* Chagal-Feferkorn, *supra* note 4, at 78. On the compensation-deterrence theory of tort law; *see generally* John C. P. Goldberg, *Twentieth-Century Tort Theory*, 91 GEO. L.J. 513, 521–37 (2002).

beneficial when improving societal outcomes or reducing harm.[95] Even if AI underperforms humans, the resultant cost savings may justify its use, leading developers and deployers to accept risks rather than impose stricter controls. Second, the analysis must account for AI's potential for improvement, not just its current performance. AI may already reduce harm compared to humans,[96] as seen in autonomous vehicles.[97] Legal frameworks should incentivize ongoing advancements—not merely the achievement of human-level performance—to foster safer, more efficient innovations.

Although more complex combinations of liability regimes are possible, this Article focuses on the primary regimes discussed above due to their academic scope.[98]

Tort law employs various liability regimes that each shape party incentives. There is one most basic yet often overlooked rule: Losses remain where they fall.[99] This rule applies by default or when other regimes do not. Though seemingly passive, this rule reflects a deliberate policy choice with significant consequences.[100] In a no-liability framework, injurers lack incentives to curb harmful behavior, prioritizing profit while victims bear the burden of avoiding harm.[101]

In contrast, strict liability removes incentives for victims but strongly incentivizes the injurer to take steps to mitigate risk and

---

95. Furthermore, if a superior AI agent offers greater benefits, the considered AI agent may become undesirable. *See infra* Section III.B.

96. *Cf.* Ryan Abbott, *The Reasonable Computer: Disrupting the Paradigm of Tort Liability*, 86 GEO. WASH. L. REV. 1, 6–7, 18 (2018).

97. *Cf.* Selbst, *supra* note 3, at 1326.

98. Examples include the adoption of a contributory negligence correction on the part of the victim, Landes & Posner, *supra* note 92, at 876, or a comparative negligence rule to inspire optimal care for both parties, Shavell, *supra* note 92, at 184. Such regimes can inspire optimal care for both victims and injurers, *id.* at 188, although some differences might persist concerning the costs of administering the regime, *id.* at 181, and the impact on activity levels, Landes & Posner, *supra* note 92, at 874–77; *cf.* Shavell, *supra* note 92, at 201–02. Moreover, any liability regime that entails claims (i.e., not a no-liability scheme) is slow and expensive. *Cf.* Chagal-Feferkorn, *supra* note 4, at 86 (discussing products liability claims).

99. *E.g.*, Shavell, *supra* note 92, at 183–84 ("no liability").

100. Landes & Posner, *supra* note 92, at 872.

101. *Id.* at 872—73.

invest in accident prevention.[102] The injurer is motivated to prevent harm when the cost of prevention is lower than the expected cost of the harm for which they would be liable.[103] This serves as a preventative mechanism, encouraging producers to take sufficient steps to ensure product safety.[104] When the cost of prevention exceeds the harm, injurers may choose to compensate victims instead.[105]

Strict liability is particularly effective when the harm occurs irrespective of the victim's level of care.[106] This rule is generally favored when the injurer is best positioned to evaluate the benefits and costs of incident prevention and is thus better equipped to adapt their behavior accordingly.[107]

A pure negligence rule, which applies solely to the injurer, offers a distinct approach. Under this rule, the injurer is incentivized to exercise a specific level of care, as they are liable if they fail to meet the required standard.[108] Negligence rules are often considered advantageous when importance is also placed on encouraging victims to adjust their own level of care and activity.[109] Once the injurer meets the standard of care, any residual harm is borne by the victim.[110] However, even if the injurer behaves negligently, a negligence rule may fail in incentivizing them to limit excessive activity levels.[111]

In addition, it is crucial to recognize that harm arising from AI agents is often bilateral or even multilateral.[112] Such harm depends not only on the level of care exercised by the deployer of the agent but also on the actions of the victim or other parties within the agent's "supply

---

102. *Id.* at 873.

103. *Id.*

104. Lai, *supra* note 4, at 618.

105. Landes & Posner, *supra* note 92, at 873.

106. *Id.*

107. This is at least implicitly part of the rationale underlying strict liability. *See, e.g.*, RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2 cmt. a (AM. L. INST. 1998) (stating that manufacturers determine the level of quality control and are thus able to predict the number of flawed products entering the marketplace).

108. Landes & Posner, *supra* note 92, at 873.

109. Shavell, *supra* note 92, at 184; Bar-Gill & Porat, *supra* note 90, at 87.

110. Shavell, *supra* note 92, at 183—84.

111. Shavell, *supra* note 92, at 197; Bar-Gill & Porat, *supra* note 90, at 87.

112. On the notion of bilateral accidents, see Shavell, *supra* note 92, at 182.

chain," such as developers.[113] Therefore, the liability regime should provide appropriate incentives to all distinct parties.[114] This Article's analysis primarily considers three key actors whose behavior influences the benefits and harms of AI systems: the developer, the deployer, and the potential victim.[115]

To cement the analysis, this Article examines three specific examples to illustrate the desirable normative AI framework:

> **Example 1:** A doctor uses an AI agent developed by an external AI developer to analyze patient scans.[116] The patients may be unaware of the AI tool's involvement. Harm arises if the AI agent misdiagnoses a scan and the doctor fails to intervene or identify this error.
>
> **Example 2:** A driver uses a semi-autonomous vehicle developed by an external AI developer. The vehicle

113. Jennifer Cobbe, Michael Veale & Jatinder Singh, *Understanding Accountability in Algorithmic Supply Chains*, *in* FACCT '23: PROCEEDINGS OF THE 2023 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1186, 1187—89 (2023).

114. Intuitively, one might expect the Coase theorem to limit the multitude of relevant actors. *See generally* R. H. Coase, *The Problem of Social Cost*, 3 J.L. & ECON. 1 (1960); Christine Jolls et al., *A Behavioral Approach to Law and Economics*, 50 STAN. L. REV. 1471, 1483 (1998); Russell B. Korobkin & Thomas S. Ulen, *Law and Behavioral Science: Removing the Rationality Assumption from Law and Economics*, 88 CAL. L. REV. 1051, 1094–95 (2000); Shavell, *supra* note 92, at 102). However, this theorem is not adequately confirmed by empirical data. Daniel Kahneman et al., *Experimental Tests of the Endowment Effect and the Coase Theorem*, 98 J. POL. ECON. 1325, 1329 (1990); *see also* Jolls et al., *supra*, at 1483; Radmilo Pešić, *Information Asymmetry and the Coase Theorem Fallacy*, 2 ERUDITIO 67, 68 (2020)). Furthermore, the application of this theorem is obstructed by the inherent unpredictability of AI systems. *Cf.* Pešić, *supra*, at 70 (in the context of information asymmetry); Paul A. Brehm & Eric Lewis, *Information Asymmetry, Trade, and Drilling: Evidence from an Oil Lease Lottery*, 52 RAND J. ECON. 496, 496 (2021) (describing information asymmetry as 'trade frictions'); Mark A. Sayre & Kyle Glover, *Machines Make Mistakes Too: Planning for AI Liability in Contracting*, 15 CASE W. RES. J.L. TECH. & INTERNET 357, 384 (2024) (implied).

115. While this simplifies the complex AI system supply chain, the analysis remains adaptable to other relevant actors.

116. *See, e.g.*, John Mongan et al., *Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers*, 2 RADIOL. A.I. 1, 1 (2020).

malfunctions, causing a collision with a cyclist who is unaware of the AI technology in the vehicle.

Example 3: A client employs an AI tool, developed by an external AI company, to replace services previously provided by their legal counsel.

Although these scenarios are more complex than traditional tort cases—featuring three actors rather than a simple injurer-victim dichotomy—the societal impact of deploying AI systems involves both their economic value to the *deployer* and the negative economic impact on the *victim*. Ideally, a tort law regime would incentivize all parties to maximize this net societal value.

All three examples highlight that, for many deployers of AI agents, relying on or deploying such technology is an active choice. While liability can arise from poor AI performance, leading to harm, a more nuanced challenge emerges when an AI agent's capabilities inspire reliance. For instance, the doctor in Example 1 might choose to ignore the AI agent's output, or the party in Example 3 might continue using a lawyer while leveraging the AI agent for supplementary input. In that way, these scenarios underline the potentially nuanced role of deployers. The tool in Example 1 might be deployed more autonomously—as a true AI agent—or with more oversight of its process or outputs, edging closer toward a more conventional AI tool. An effective liability regime should seamlessly and coherently accommodate this spectrum of deployment, covering both instances where AI agents operate autonomously with minimal oversight and where they function purely as assistive tools under human control.

In scenarios occurring along the spectrum of deployment, the core issue lies in the consequences of relying on the AI agent; this reliance creates unique harm scenarios. Examples 1 and 3 demonstrate that reliance is often voluntary, while Examples 1 and 2 show that some parties, such as the patient or cyclist, may not have a choice. An effective liability regime must address both voluntary and involuntary reliance on AI agents to ensure fair and optimal outcomes and incentivize the relevant parties.

A. Liability for Voluntary Reliance on AI Agents

First, consider cases where an AI deployer voluntarily relies on the system and, as a result, experiences harm due to its performance or decisions. This is exemplified by the client in Example 3 and the doctor in Example 1. For simplicity, this Article assumes that the roles of software developers and other providers higher up the AI supply chain are combined into a single entity, referred to as the "developer" of the AI agent.[117] Yet, if these roles are distinct, their relative liability would still primarily be determined by the same principles outlined here.[118]

AI agents are often chosen because they offer improved outcomes or similar results with greater efficiency.[119] Since the deployer voluntarily decided to employ the system, it is reasonable to assume that they perceived some benefit. For instance, in Example 1, the doctor could have continued treating patients without the AI agent but actively chose to deploy it to gain an advantage.

Assuming the developer has delivered the AI agent *as is*—regardless of whether they could have adhered to a higher standard of care to improve it—the agent's impact depends on the level of care exercised by the voluntary deployer.[120] The deployer's active choice underscores their role in managing the risks of deploying the AI agent.

While the voluntary deployer of an AI agent chooses to deploy it with the expectation of a net benefit, the inherent unpredictability of AI agents makes it practically impossible to eliminate damage entirely. For example, damage could theoretically be avoided by manually verifying every piece of AI output. This, however, would significantly

---

117. Similarly, this Article uses the term "deployer" broadly and naturally, without being restricted to, for example, the definition given in the European AI Act, *supra* note 13, art. 3(4).

118. In many instances they may be treated as a single actor by virtue of the Coase theorem, *see* Coase, *supra* note 114.

119. *Supra* Section II.B.

120. The system's (negative) impact is influenced both by its development (which was fixed here) and its deployment. For instance, a deployer could independently verify each system output, preventing any (new) harm regardless of the system's quality—though this would largely negate its economic benefits. Conversely, the deployer could blindly rely on the system for critical decisions. In both cases, the way in which the system is deployed clearly impacts the relevant risks.

diminish the AI agent's utility or require excessive care from the developer, making the agent unattractive for deployment.

If the agent's use harms the voluntary deployer, several key points emerge. First, under a no-liability regime, the voluntary deployer is incentivized to adjust their behavior and level of care to avoid damage.[121] This adjustment protects the deployer and also contributes to maximizing societal benefits overall.

The risk of harm does not depend solely on the AI system's performance. If a tool performs poorly, deployers can mitigate harm by exercising caution when relying on it.[122] In scenarios where a voluntary deployer's unwarranted reliance on the AI agent results in harm, a no-liability rule naturally incentivizes deployers to carefully evaluate and adjust their reliance, encouraging responsible behavior.

Second, instances of harm in voluntary deployment underscore the importance of ensuring that AI deployers can make informed decisions when relying on their systems. An ideal liability regime should require AI developers to offer clear and accurate information, enabling deployers to evaluate whether to rely on the system. More specifically, developers should highlight relevant features that help deployers assess the system's suitability for their needs.

In an ideal scenario, system developers would only be incentivized to give voluntary deployers all necessary information for informed decision-making. If harm to external parties is excluded from consideration—which this Article discusses below[123]—and only the harm to voluntary deployers is considered, system developers would not necessarily require additional incentives to improve the AI agent's performance.[124] Market dynamics could address this, as informed

121. Landes & Posner, *supra* note 92, at 872–73 (with the victim as the deployer).

122. *See also* Shavell, *supra* note 92, at 182 (discussing more "traditional" contexts without AI and the general possibility for victims to impact losses).

123. *See infra* Section III.B.

124. This is based on the assumption that a truly "voluntary" deployer is well-informed regarding the relevant system risks and can thus be expected to make rational choices regarding those risks. *See generally* Coase, *supra* note 114 (discussing the Coase theorem).

deployers would naturally select tools that best match their needs and favor those with better performance.[125]

However, providing perfect information is inherently challenging. This difficulty is not unique to AI. For instance, vehicles like the one in Example 2 may malfunction in various ways, regardless of AI involvement. AI complicates matters further due to its inherent unpredictability.[126] If the developer cannot fully predict the system's behavior, thoroughly informing deployers becomes nearly impossible. This information gap creates challenges for market efficiency, as deployers may lack the data needed to make optimal choices, potentially distorting free-market dynamics.[127]

As a result, the liability regime should include mechanisms that incentivize parties to achieve societally desirable outcomes. Specifically, liability law should encourage AI developers to enhance system performance and minimize both the likelihood and extent of harm. The importance of this "performative" aspect of the liability regime grows with the increasing complexity of AI agents. In more complex environments, achieving optimal societal outcomes through free-market mechanisms becomes significantly harder.[128] For highly complex applications, such as autonomous vehicles, a strict liability regime is likely the most appropriate approach.

Additionally, there is value in incentivizing AI developers to improve system oversight by voluntary deployers, such as by ensuring the system is explainable. Enhanced explainability can help reduce the harm incurred by victims for a given level of care, reinforcing the broader objectives of the liability framework.[129]

Two additional observations refine this conclusion. First, voluntary deployers of AI agents may externalize some of the harm

---

125. *See supra* note 114.

126. *Cf.* Pešić, *supra* note 114 (in the more general context of information asymmetry).

127. This is in line with the observation that the Coase theorem likely does not fully apply in this context. *Supra* note 114.

128. *See supra* note 114.

129. Explainability facilitates human oversight. *Supra* Section I.B. Explainability can thus enhance the effectiveness and impact of some given form of oversight, corresponding to a certain level of care.

caused by the AI agents.[130] For instance, in Example 1, the patient, rather than the doctor, bears the primary harm if the doctor incorrectly relies on the AI agent. This highlights the importance of the deployer's potential liability toward third parties, such as patients, which this Article addresses below.[131] It also underscores the need to incentivize deployers to minimize external harm, irrespective of the AI agent's performance. Deployers should, therefore, be precluded from holding the system developer liable if the harm is due to their own carelessness.

Second, AI agents may externalize benefits as well.[132] For example, if AI agents improve medical decision-making, they can enable doctors to provide better and more efficient patient care, as in Example 1. This has significant implications, which this Article explores further below.[133]

As a result, the unpredictability and complexity of AI agents often make it difficult or impossible for voluntary deployers to fully grasp the system's potential dangers and limitations. Liability law must, therefore, create incentives for all parties to achieve socially desirable outcomes. System developers should be incentivized to inform deployers comprehensively and to optimize system performance. At first glance, a strict liability regime appears effective for achieving these goals.[134] It would also promote the adoption of AI agents, as voluntary deployers could obtain compensation for any damage incurred during system use.[135]

However, the voluntary developer's liability should decrease in proportion to the deployer's responsibility. A contributory or comparative negligence rule would help achieve this balance.[136] The level of care expected from voluntary deployers should increase as

130. Externalization denotes that this harm is not immediately experienced by the relevant actor(s). *See also* Lai, *supra* note 4, at 619.

131. *See infra* Section III.B.

132. *See, e.g.*, Kyle Colonna, *Autonomous Cars and Tort Liability*, 4 J.L. TECH. & INTERNET 81, 111–12 (2012) (autonomous vehicles cause fewer accidents).

133. *See infra* Section III.B.

134. *See also* Anat Lior, *supra* note 3.

135. Landes & Posner, *supra* note 92, at 876–77.

136. *See generally* Shavell, *supra* note 92, at 186–87 (discussing the economic implications of contributory and comparative negligence).

circumstances and the information provided by the agent developer allow them to make more informed choices.

B. External Liability of the Developer and Deployer

The case for imposing liability on voluntary AI deployers and developers is even stronger when the victim of AI-related harm did not choose to interact with the system, as exemplified by the patient in Example 1 and the cyclist in Example 2. In traditional cases, one might argue that placing harm on the victim incentivizes them to avoid harmful interactions, effectively making them feel the economic damage caused by the system.[137] However, this reasoning does not apply when victims cannot reasonably avoid harm due to their unawareness of the AI system's presence.[138]

In such situations, maximizing the overall balance of benefits and harms necessitates shifting responsibility to the developer and deployer of the system. These parties, rather than the unaware victim, should adjust their level of care. This adjustment ensures that developers and deployers experience the full benefits and harms their systems create.[139] The argument becomes particularly compelling when the system causes net societal harm—where the harms outweigh the benefits—but it is equally relevant in other contexts.

At first glance, these circumstances justify adopting a strict liability regime. Developers and deployers are typically best positioned to assess the societal costs and benefits associated with their systems, making it logical to focus incentives on their behavior.[140]

However, two factors complicate this scenario. First, as noted previously, AI systems may externalize costs along with benefits.[141] While cost externalization is less significant for the victim in these cases, as they typically bear most of the harm, the externalization of benefits remains a significant issue. For instance, in Example 2, the

137. Bar-Gill & Porat, *supra* note 90, at 88.

138. *See also* Shavell, *supra* note 92, at 186–87 (explaining how it is useless to (additionally) incentivize victims that have no control over the harm).

139. *See generally* Shavell, *supra* note 92, at 188 (describing how the ideal liability rule depends on the level of care it seeks to incentivize).

140. *See also* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2 cmt. a (AM L. INST. 1998) (motivating the rationale behind strict liability).

141. *Cf.* Lai, *supra* note 4, at 618.

developer of a semi-autonomous vehicle may produce a high-performing system that generates societal benefits, such as reducing overall accident rates. However, these benefits may not directly aid the developer or deployer, as they accrue broadly to society or other parties.[142]

This illustrates how benefits can be distributed unevenly. For instance, a semi-autonomous vehicle's societal advantage in reducing accidents may only interest the deployer when it minimizes damage to their own vehicle or liability exposure. Developers and deployers may be unwilling to pay for these broader benefits.[143]

The societal benefit of fewer car accidents is not necessarily confined to individual instances. Even if an AI malfunction causes an accident, the incident may contribute to the overall safety of future vehicles by allowing the system developer to improve the technology. These considerations suggest that the liability regime should not automatically allocate the harm incurred by the agent deployer to the developer. Similar reasoning applies to the doctor in Example 1, who might improve their ability to treat future patients through the AI system or agent, even though the victim of AI-related harm may not recognize or value such indirect benefits.

However, benefit externalization is not unique to autonomous vehicles and is also present in other AI applications. For instance, AI tools aiding innovative drug research may benefit future generations of patients, even as today's patients bear the consequences of the tool's error rate.[144] Therefore, it may not be socially desirable for the developer and deployer to bear the full societal impact of the AI agent—the combination of benefits and harms—but rather a proportionally reduced incentive. Externalized benefits might create hidden societal gains that either party does not directly experience.[145]

142. *See, e.g.*, Keith N. Hylton, *A Positive Theory of Strict Liability*, 4 REV. L. & ECON. 153, 156–57 (2008) (discussing cross-externalization).

143. *Cf.* Landes & Posner, *supra* note 92, at 872 ("Why should B expend any resources on care when the benefits of that expenditure will inure to A?").

144. *See, e.g.*, Lai, *supra* note 4, at 618–19.

145. *See also* Bar-Gill & Porat, *supra* note 90, at 92 (discussing benefits experienced by third parties).

Imposing the total economic impact of harm on developers and deployers could discourage investment in and development of AI systems and agents,[146] thereby undermining the societal advantages of these externalized benefits.[147] Instead, developers and deployers should face a reduced liability, accounting for the external benefits of the AI agent. This approach balances the need to incentivize continued AI development with the broader societal interest.

However, reducing the liability of developers and deployers would shift some of the harm to victims, thereby incentivizing them to alter their behavior—an outcome previously deemed undesirable.[148] This trade-off highlights the complexity of crafting an optimal liability regime for AI agents, where both direct and indirect effects must be carefully weighed.

Such external benefits could support an argument for introducing external corrections to liability law. In critical and complex domains, AI developers and deployers may not be able to capitalize on the broader societal benefits their systems generate. Ideally, society should bear the cost of such benefits, rather than placing the entire burden on developers or deployers. This approach could involve allocating part of the harm to societal mechanisms, such as mandatory insurance schemes or government-funded subsidies, which, when combined with a strict liability regime, can ensure a balanced distribution of responsibilities.[149]

A second, related complexity concerns the threshold for determining harm. Harm should not only be assessed by comparing the situation to one without an AI agent but also to the outcomes reasonably achievable with AI systems and agents. In some cases, there may be no harm to the victim compared to a scenario without AI. However, liability remains essential to incentivize developers and

146. *Id.* at 618; *cf.* Bryant Walker Smith, *Automated Driving and Product Liability*, 2017 MICH. ST. L. REV. 1, 2 (2017) (discussing products liability); Selbst, *supra* note 3, at 1322 (describing this statement rather than defending it).

147. *Cf.* Colonna, *supra* note 132 (implicitly); Chagal-Feferkorn, *supra* note 4, at 102; Lai, *supra* note 4, at 598–99. On the impact of products liability, *see generally* Lai, *supra* note 4, at 618–19; A. Mitchell Polinsky & Steven Shavell, *The Uneasy Case for Product Liability*, 123 HARV. L. REV. 1437, 1473 (2010).

148. *See supra* notes 141–143 and accompanying text.

149. For this approach, see Funkhouser, *supra* note 93, at 458.

deployers to maximize net societal benefits, particularly those impacted by the agent's use. The mere presence of AI in the market alters the expectations and capabilities of its voluntary deployers, effectively externalizing certain benefits to victims and raising the standard of care for existing service providers.[150]

Example 1 illustrates this well. If the patient receives care equivalent to what they would have had without the AI agent, this does not negate the need for incentives to improve care. Comparable treatment might indicate responsible agent development, but is insufficient to conclude that no liability incentives are necessary for the agent deployer.[151] Liability remains crucial to ensure that developers and deployers continuously strive to enhance their AI agents, maximizing societal and individual benefits.

Similarly, the fact that a developed AI agent outperforms humans does not eliminate the need to incentivize developers to further improve the agent. It is desirable to encourage developers to make their agents as effective and reliable as reasonably possible, as that would maximize societal benefits.[152]

From a public policy perspective, it is undesirable for victims of AI-related harm—whether due to inadequate AI use or underperformance when more could reasonably have been done—to bear that harm. Such an outcome provides no meaningful incentive, as victims are generally not able to alter their behavior in response. Consequently, all harm should be allocated to the developer and deployer, except for certain cases where it is more appropriate for harm to be distributed across society due to benefit externalization.[153] A strict liability regime is, therefore, necessary to incentivize developers and deployers to optimize their agents and deployment. This approach ensures that victims of harm are not incentivized inappropriately.[154]

---

150. *Cf.* Griffin, *supra* note 39, at 98.

151. *See supra* note 147.

152. For instance, the doctor in Example 1 may still act negligently by deploying an AI agent that, despite generally outperforming human doctors, is outdated and inferior to other reasonably accessible AI systems.

153. *Supra* note 147.

154. *See, e.g.*, Bar-Gill & Porat, *supra* note 90, at 95.

Regarding the relative liability of AI agent developers and deployers, the principles outlined earlier in Section B remain applicable. The behavioral test for determining whether the developer might escape liability should become more stringent as their knowledge of the agent and its limitations improves. The more informed their choice to rely on the agent, the greater their responsibility to ensure its safety and effectiveness.[155]

C. Holistic Framework

Bringing together all elements of this analysis, involuntary victims of AI-related harm should have a liability claim against both the developer and the voluntary deployer of the AI agent.[156] Liability should be shared to provide both parties—AI developers and deployers—with sufficient incentives to exercise appropriate levels of care.[157]

The relative responsibility of deployers should increase with their understanding of the AI agent and its limitations, incentivizing developers to inform and educate them.[158] Developers also have a clear incentive to improve the agent's explainability, as it reduces the risk of harm—and potential liability—for a given level of care exercised by the deployer.

This Article's analysis suggests that the deployer should not be subject to strict liability, as that regime lacks the necessary flexibility. Instead, the deployer's decision to rely on the AI agent should be assessed based on the information provided by the developer and the circumstances, such as the agent's complexity and its domain of application. As the deployer becomes better equipped to make an informed decision, their duty of care should rise accordingly.[159] In simple cases where the deployer has full access to relevant information, their ideal liability regime may closely resemble strict liability.[160]

---

155. *See supra* Section III.A.

156. *See, e.g.*, Smith & Fotheringham, *supra* note 93, at 143 (discussing negligence liability).

157. *Cf. id.* at 143 (discussing negligence liability).

158. *Supra* Section III.A.

159. *Supra* Section III.A.

160. *See generally* Abbott, *supra* note 96, at 35 (discussing how negligence liability approaches strict liability if the standard of care is increased).

Given the externalization of some benefits, it may be desirable to offset part of the harm through mechanisms such as insurance schemes or government-funded subsidies.[161] This externalization also contributes to the "reasonableness" of using AI agents when evaluated under a standard of negligence.[162]

Interestingly, this analysis closely parallels evaluations of liability for traditional tools and products that do not involve AI. To a large extent, the key attributes a liability regime should embody to effectively regulate AI agents remain "traditional," which is unsurprising given that many characteristics of AI agents are not unique from an economic perspective.[163] The primary challenges posed by AI agents stem from their complexity and the externalization of benefits, both of which are also observed in certain traditional domains.

IV. DESCRIPTIVE ANALYSIS OF AI LIABILITY

A. General

Having outlined a normative framework for AI agent liability, this Article now turns to the existing legal regimes to assess their alignment with the criteria established by the normative analysis. This discussion focuses on two primary regimes: negligence liability and products liability.[164]

Although negligence liability is considered the "general" or default regime, it is often[165] overlooked in the context of AI systems and agents.[166] This is largely due to the emphasis on programmers—i.e.,

161. Some authors have proposed such a "no-fault" scheme, for example in the form of insurance, in the context of autonomous vehicles. *See* Funkhouser, *supra* note 93, at 458.

162. *Cf.* Abbott, *supra* note 96, at 39–40 (in the context of autonomous vehicles).

163. The preceding normative analysis helped nuance the distinction between AI-related and "traditional" non-AI products. Consequently, an optimal legal framework should not differ significantly from the "traditional" framework either.

164. *See generally* Henson, *supra* note 4 (discussing the potential applicability of strict liability for abnormally dangerous activities).

165. *See, e.g.*, Moffatt v. Air Canada, 2024 CanLII 149, paras. 24–32 (Can. B.C. C.R.T.) (discussing negligent misrepresentation).

166. *Cf.* Selbst, *supra* note 3, at 1318.

system developers[167]—and the corresponding reliance on products liability.[168] In the European Union, the growing prevalence of AI systems and agents has been a significant impetus for updating its products liability framework.[169]

The preference for a products liability regime is often attributed to concerns about the lack of explainability and predictability of AI agents, which are seen as challenging for negligence law.[170] Unlike negligence liability,[171] products liability does not necessarily hinge on a breach of duty of care.[172] Some aspects of products liability incorporate strict or fault-free liability,[173] which is argued to provide economic benefits by enabling producers to better distribute liability costs or obtain insurance coverage.[174]

If negligence liability is addressed at all in the context of AI systems and agents, the discussion is often narrowly focused—for instance, on scenarios where the agent deployer acts negligently by failing to adhere to the system user manual.[175] It is important to emphasize, however, that the potential liability of the developer under products liability does not preclude the relevance of negligence liability.

To establish negligence liability, plaintiffs must demonstrate the existence of a duty of care, a breach of that duty,[176] the harm suffered

167. *Id.* at 1328.

168. *Cf. id.* at 1322.

169. *See Proposal for a Directive of the European Parliament and of the Council on Liability for Defective Products*, COM (2022) 495 final (Sept. 28, 2022).

170. *Infra* Section IV.C.1.

171. *Infra* Section IV.C.1.

172. *Infra* Section IV.C.2.

173. *E.g.*, Abbott, *supra* note 96, at 13–14.

174. *See, e.g.*, Escola v. Coca-Cola Bottling Co. of Fresno, 150 P.2d 436, 441 (Cal. 1944) (Traynor, J., concurring); *cf.* John W. Wade, *On the Nature of Strict Tort Liability for Products*, 44 MISS. L.J. 825, 837–38 (1973) (incorporating the feasibility of carrying insurance by the product developer in the analysis of whether a product is defective); Chagal-Feferkorn, *supra* note 4, at 78.

175. Gary E. Marchant & Rachel A. Lindor, *The Coming Collision Between Autonomous Vehicles and the Liability System*, 52 SANTA CLARA L. REV. 1321, 1327 (2012) (discussing fully autonomous vehicles); *cf.* Selbst, *supra* note 3, at 1328–29.

176. Selbst, *supra* note 3, at 1330.

by the plaintiff, and a causal link between the breach and the harm.[177] Products liability, in contrast, requires[178] that the AI system or agent qualifies as a product,[179] that the developer sold the agent,[180] and that the agent in its unaltered state[181] caused damage to a party[182] due to[183] a defect in the agent.[184] Each of these elements presents unique challenges in the context of AI.

This analysis does not focus extensively on the damage requirement under either regime, as it remains largely unaffected by the presence of AI agents.[185] However, it is worth noting that negligence liability sometimes limits recovery for purely economic loss,[186] a restriction consistently applied in products liability.[187]

For clarity, this Article compares both regimes and explicitly identifies the liable party when not apparent. First, however, it examines the general applicability of products liability to AI agents.

B. AI Agents as Products?

A fundamental requirement of the products liability regime is that it applies exclusively to products.[188] However, this criterion may not

177. *See, e.g.*, Gerstner, *supra* note 3, at 246; Michael D. Scott, *Tort Liability for Vendors of Insecure Software: Has the Time Finally Come*, 67 MD. L. REV. 425, 442 (2008).

178. Lai, *supra* note 4, at 613.

179. Chagal-Feferkorn, *supra* note 4, at 83; Lai, *supra* note 4, at 613; *see also* Gerstner, *supra* note 3, at 250 (implicitly).

180. Lai, *supra* note 4, at 613.

181. *Id.*

182. Lai, *supra* note 4, at 613.

183. Griffin, *supra* note 39, at 79; Lai, *supra* note 4, at 613.

184. Lai, *supra* note 4, at 613. *Cf.* Gerstner, *supra* note 3, at 250.

185. The types of harm recognized by law remain unchanged, regardless of whether an AI system was involved in breaching a relevant liability standard.

186. *See generally* RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR ECONOMIC HARM § 1 (AM. L. INST. 2012) (stating that there is no general duty of care to avoid economic loss); Danielle Sawaya, *Not Just for Products Liability: Applying the Economic Loss Rule beyond Its Origins*, 83 FORDHAM L. REV. 1073, 1077 (2014).

187. For products liability, *see* RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 1 (AM. L. INST. 1998); Sawaya, *supra* note 186, at 1077.

188. *See* sources cited *supra* note 179.

be straightforwardly met for AI systems and agents.[189] Traditionally, products are understood as necessarily tangible.[190] Similarly, certain types of traditional, information-providing software have at times been classified as services rather than products, given their functional similarity to services rendered by human professionals.[191] This reasoning could extend to AI agents, potentially leading to their classification as services—particularly in the case of information-providing AI.[192] Such challenges have prompted the European Union to revisit its products liability directive to explicitly encompass AI systems.[193]

The argument that an AI system constitutes a service is particularly compelling when it is deployed by its manufacturer, who provides outputs in exchange for compensation.[194] In such scenarios, the agent might also fail the second requirement of the products liability regime—that it be "sold."[195] Nevertheless, the requirement for a physical presence has generally been interpreted broadly[196] to encompass software so long as it has a tangible embodiment accessible to the user.[197]

Given this broad interpretation, it is likely that courts will classify AI agents as products for products liability.[198] The Third Restatement

189. Chagal-Feferkorn, *supra* note 4, at 82. The Author, however, does not take this position. *Cf.* Gerstner, *supra* note 3, at 250. For medical contexts, compare Mousa Alshanteer, *A Current Regime of Uncertainty: Improving Assessments of Liability for Damages Caused by Artificial Intelligence*, 21 N.C. J.L. & TECH. 27, 42 (2019); Gary E. Marchant & Lucille M. Tournas, *AI Health Care Liability: From Research Trials to Court Trials*, 18 J. HEALTH & LIFE SCI. L. 25, 32 (2019).

190. *Cf.* Scott, *supra* note 177, at 434–35 (discussing software more generally); Chagal-Feferkorn, *supra* note 4, at 82.

191. Scott, *supra* note 177, at 461; Chagal-Feferkorn, *supra* note 4, at 82–83; Lai, *supra* note 4, at 614; *see also* Gerstner, *supra* note 3, at 250.

192. *See also* Chagal-Feferkorn, *supra* note 4, at 82.

193. *See* Proposal for a Directive on Liability for Defective Products, *supra* note 169.

194. George S. Cole, *Tort Liability for Artificial Intelligence and Expert Systems*, 10 COMPUT. L.J. 127, 160 n.109 (1990).

195. As there is no sale in such scenarios, the developer of the AI may, however, be liable on the basis of services liability. *Id.*

196. *Id.* at 160; *cf.* Gerstner, *supra* note 3, at 251.

197. Cole, *supra* note 194, at 160.

198. Scott, *supra* note 177, at 466–67; Chagal-Feferkorn, *supra* note 4, at 83–84.

of Torts on Products Liability supports this view, stating that AI systems could qualify as products if (1) "justice [is served by] imposing the loss on the manufacturer who created the risk and reaped the profit" and (2) there is "difficulty in requiring the injured party to trace back along the channel of trade to the source of the defect to prove negligence." [199]

This clarification suggests that the classification of AI agents as "products" should not deter further examination of products liability implications. At the same time, negligence liability remains relevant, as the Third Restatement of Torts favors it where feasible. [200]

C. *Liability Standard*

1. *Negligence Liability*

a. *General*

Negligence liability hinges on a breach of duty of care. [201] This duty of care can be addressed briefly, with two primary duties typically recognized for software developers: (1) the duty to develop reliable software and (2) the duty to provide users with adequate instructions for its reliable use. [202] These duties are analogous to those owed by sellers of products to their customers. [203]

However, the presence of a relevant duty of care is not always guaranteed for AI agent deployers. [204] This Article focuses on contexts where a duty of care exists, allowing an analysis of potential breaches when relying on AI agents. Such duties are common in domains where AI is widely deployed.

---

199. RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 19 reporter's notes to cmt. a; *cf.* Trent v. Brasch Mfg. Co., Inc., 477 Ill. App. 3d 586, 590 (1985).

200. *See* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 19 reporter's notes to cmt. a.

201. *E.g.*, Gerstner, *supra* note 3, at 246–47; Scott, *supra* note 177, at 442–43.

202. Scott, *supra* note 177, at 443.

203. Gerstner, *supra* note 3, at 246–47.

204. Selbst, *supra* note 3, at 1319 (discussing decision-assistance by AI systems in employment, lending, retail, policing and agriculture).

For example, in the case of autonomous vehicles, the duty may involve ensuring the safety of others in traffic.[205] In healthcare, where AI agents can support or autonomously carry out medical decision-making, the duty of care could be the one owed by medical practitioners to their patients.[206] Similarly, if an AI agent is used to inform a counterparty through a chatbot, the deployer must ensure that it does not result in negligent misrepresentation.[207]

b. Breach of Duty of Care by Developers

For developers, the key challenge is not the existence of a duty of care but understanding the implications of their AI agent's deployment, especially when that use results in harm. Developers are expected to ensure that the systems they create are "up to standards,"[208] though the precise meaning of this remains somewhat abstract. As professionals,[209] AI developers are generally held to a high standard of care, even though courts have historically hesitated to impose such a standard on system programmers.[210]

Some argue that the unpredictability of AI systems and agent behavior complicates the application of negligence law.[211] However, this reasoning is less convincing for current-generation AI systems, as this Article discusses more elaborately below when considering deployers.[212]

Several existing criteria can help demonstrate that a developer exercised due diligence in creating an AI agent. The foremost consideration is evaluating how well the agent performs for its

205. *E.g.*, STUART M. SPEISER ET AL., AMERICAN LAW OF TORTS § 9:37 (2024) (discussing the duty of care with regard to passengers).

206. *Cf.* Smith & Fotheringham, *supra* note 93, at 133–34 (discussing English common law).

207. *See* Moffatt v. Air Can., 2024 CanLII 149, paras. 24–32 (Can. B.C. C.R.T.).

208. *Cf.* AI Act, *supra* note 13, art. 15 (establishing this requirement in the European Artificial Intelligence Act).

209. Gerstner, *supra* note 3, at 247 (stating that professionals are individuals who have a specific level of knowledge and ability, and who use that knowledge and ability to take on work that requires special skill).

210. *Id.*

211. *See, e.g.*, Smart et al., *supra* note 44, at 43–44.

212. *Infra* Section IV.C.1.

intended use.[213] While this may seem straightforward, it is far from simple. There are numerous ways to measure an AI agent's performance, as is readily apparent for classification systems, such as tools that analyze patient scans to diagnose tumors.[214] Statistical measures like type I and type II errors can be used to assess performance.[215] A type I error refers to a false positive—for example, a healthy patient's scan being identified as problematic. A type II error refers to a false negative, such as a troubling scan being misclassified as healthy.

Assessing AI agent performance is, however, challenging. One factor is that error significance varies by context.[216] This is particularly evident in medical classification tools, where type II errors can be more severe.[217] A type I error may lead to unnecessary follow-ups, reducing efficiency, while a type II error—failing to flag a concerning scan—can cause serious harm. However, excessive false positives can also render the agent impractical.[218]

A key challenge is the lack of a single metric for evaluating system accuracy. While an AI agent that consistently outperforms another is

213. In line with the duty to develop reliable software, see *supra* note 202 and accompanying text.

214. Mongan et al., *supra* note 116 and accompanying text.

215. *See, e.g.*, CHRISTIAN HEUMANN ET AL., INTRODUCTION TO STATISTICS AND DATA ANALYSIS 213 (2016).

216. Much like in "traditional" negligence contexts, the care that should be taken is thus dependent on the potential damage that could arise, as reflected in the Hand formula. *See* United States v. Carroll Towing Co., 159 F.2d 169, 173 (2d Cir. 1947); *cf.* RESTATEMENT (THIRD) OF TORTS: PHYSICAL AND EMOTIONAL HARM § 3 (AM. L. INST. 2010). *See generally* Barbara Ann White, *Risk-Utility Analysis and the Learned Hand Formula: A Hand That Helps or a Hand That Hides?*, 32 ARIZ. L. REV. 77 (1990); Keith N. Hylton, *Information and Causation in Tort Law: Generalizing the Learned Hand Test for Causation Cases*, 7 J. TORT L. 35, 35 (2014); Daniel P. O'Gorman, *Contract Law and the Hand Formula*, 75 LA. L. REV. 127, 156 (2014).

217. Phrased differently, they typically present a more significant L in the Hand formula. *See Carroll Towing*, 159 F.2d at 173; *cf.* RESTATEMENT (THIRD) OF TORTS: PHYSICAL AND EMOTIONAL HARM § 3. *See generally* White, *supra* note 216; Hylton, *supra* note 216; O'Gorman, *supra* note 216.

218. This requires increased human supervision, diminishing the efficiency gains of deploying the AI agent.

superior, such cases may be rare.[219] Accuracy measures are not universal, and their appropriateness depends on the specific agent and its application context. Interestingly, the system model inherently defines what it means to be "better." The choice of a performance evaluation scale is crucial, as it directs training and determines whether the system meets the required standards.[220] AI systems often rely on metrics for type I and type II errors, balanced to reflect specific priorities.[221]

Another consideration is the extent of testing required to evaluate system accuracy.[222] While optimizing performance is prudent, the question remains: When is a system "good enough?" AI systems are inherently imperfect,[223] and while further training can enhance performance, limitless improvement is neither feasible nor reasonable.[224] At some point, additional training costs outweigh the benefits in time, data, and resources.[225]

---

219. It is more likely that some AI system performs better for a specific subset of use-cases while others may perform better for a (smaller) subset of different use-cases. The same point is illustrated by analyses of Large Language Models that highlight how various models may perform relatively better for some identified tasks. *E.g.*, *LLM Leaderboard - Comparison of GPT-4o, Llama 3, Mistral, Gemini and over 30 Models*, ARTIFICIAL ANALYSIS, https://artificialanalysis.ai/leaderboards/models [https://perma.cc/LEA4-P2DS] (last visited Apr. 3, 2025).

220. Metrics like the F-score can be used during system training to help the system assess whether it should implement some change to improve its performance, *see, e.g.*, REBALA ET AL., *supra* note 27, at 61.

221. For instance, "recall" measures the proportion of problematic scans correctly identified out of all problematic scans, including those missed (Type II errors). "Precision," on the other hand, evaluates the proportion of correctly identified problematic scans out of all scans flagged as problematic, including false positives (Type I errors). These metrics are complementary and are often combined into a single performance measure, such as the F-score. However, these measures are neither unique nor universally applicable. For medical tools, for example, it is generally more appropriate to prioritize reducing Type II errors, which could justify assigning greater weight to recall. *See id.* at 60–62.

222. *Cf.* Gerstner, *supra* note 3, at 248 (discussing vendors of software).

223. *Supra* note 30 and accompanying text.

224. *Supra* note 41 and accompanying text.

225. *Supra* note 42 and accompanying text.

Defining sufficient performance is challenging. A common benchmark entails surpassing human capabilities, yet proving such superiority is complex. While system performance is generally more quantifiable,[226] human performance may be harder to measure in the absence of clear metrics,[227] underscoring the importance of data in both training and evaluation. In addition, merely matching or slightly exceeding human performance may no longer suffice as AI advances. The evolving standard for AI development raises questions about responsibility. An agent's failure to outperform humans or competitors does not imply negligence.[228] It may still offer benefits such as efficiency or accessibility, shifting responsibility from developers to deployers. In such cases, optimization of deployment, rather than endless system refinement, becomes the priority.

Ultimately, evaluating AI system performance requires balancing technical capabilities with user expectations. The relevant question is not whether the agent or system delivers "absolute good" but whether it facilitates meaningful value creation through its deployment.

This highlights a second key duty of developers, as emphasized in this Article's normative analysis: equipping deployers with clear information and instructions for reliable agent use.[229] For AI agents, this requirement is particularly complex, as all AI agents can potentially produce erroneous outputs.[230] Developers must clearly communicate this possibility to deployers, ensuring they are equipped to use the agent in a way that creates value. This aligns with the

---

226. For machine learning systems, for example, system training inherently imposes some form of performance quantification to help determine whether the system should adopt some alteration. *Cf. supra* note 220 and accompanying text.

227. Even in contexts where performance is quantifiable (e.g., the number of correctly diagnosed patient scans by a doctor), such information is not always recorded.

228. Similarly, negligence is determined not by comparing to some more diligent individual but by measuring against the standard of care expected from a reasonable person. *E.g.*, SPEISER ET AL., *supra* note 205, § 9:4.

229. Scott, *supra* note 177, at 443; *cf.* RESTATEMENT (THIRD) OF TORTS: PHYSICAL AND EMOTIONAL HARM § 18 (AM. L. INST. 2012) (discussing a duty to warn people subjected to some risks); Smart et al., *supra* note 44, at 45 (discussing a duty to educate).

230. On the inherent imperfection of AI, see *supra* note 30 and accompanying text.

"crashworthiness" standard proposed for cars, which some advocate extending to computer systems.[231] If an AI agent is imperfect, developers should take steps to mitigate the adverse impact of errors, including warning deployers against uncritical reliance on the agent.[232]

In this context, error prevention may necessitate limiting the agent's autonomy by incorporating a human in the loop. For instance, most people would prefer a doctor using an AI tool to assist in treatment decisions over an AI tool that autonomously initiates treatment. Similarly, one might mandate a human driver for an autonomous vehicle.[233] This approach is especially critical when erroneous outputs could result in significant economic, moral, or health-related harm.

However, this perspective lacks nuance. As previously discussed, accepting some errors may be economically justifiable if the AI agent's overall benefit outweighs its potential drawbacks.[234] This is not a binary issue; AI agents can permit varying degrees of supervision, whether general or individualized, depending on the context and the potential impact of errors.[235]

If the agent allows for human supervision, this process is greatly enhanced by a degree of explainability.[236] An AI agent that offers explainability is inherently more reasonable than one that does not.[237]

---

231. Choi, *supra* note 30, at 86.

232. *Cf.* Smart et al., *supra* note 44, at 39 (discussing an example of a person whose hair was sucked into their autonomous vacuum cleaner as illustrating the importance of informing the victim of the system, although this Article would argue that the person who deploys the vacuum cleaner is its deployer).

233. This is the most common requirement for today's autonomous vehicle regulation in state law. *See* Atilla Kasap, *States' Approaches to Autonomous Vehicle Technology in Light of Federal Law*, 19 OHIO ST. TECH. L.J. 315, 342 (2023).

234. As in the case of benefit externalization. *See supra* note 147 and accompanying text. Furthermore, the law does not generally demand perfection from humans either. *See infra* note 257 and accompanying text.

235. This is also impacted by their level of explainability. *See supra* Section II.B.

236. *Supra* Section II.B.

237. For a given level of care—or a set amount of time and effort dedicated to supervision—an explainable AI system is expected to cause less harm than a non-explainable one, as explainability enhances oversight. Consequently, for any fixed level of care, an explainable system is likely to reduce harm

footnote continued on next page

The level of detail provided should correspond to the expertise of the average deployer. This principle aligns with the European Union's AI Act, [238] which requires AI systems to be explainable where appropriate and to support human oversight. [239]

Additionally, demonstrating diligent AI development requires developers to maintain accurate logs of the agent's performance. [240] This is especially critical for agents who continue learning during deployment or undergo updates.

c. *Breach of Duty of Care by Deployers*

The deployer's duty of care depends on their specific relationship to the victim rather than arising from a distinct obligation as an AI deployer or programmer. Similarly, the applicable standard of care is determined by a general assessment of whether an existing duty has been breached. [241] For instance, in the case of a physician, the key question is whether a reasonably competent physician would have acted similarly under comparable circumstances. [242]

Additionally, there is a general duty to exercise reasonable care when there is a risk of physical harm to others. [243] The use of an AI agent does not elevate the required standard of diligence. [244]

Some authors argue that the lack of explainability in AI systems and agents makes it impossible for deployers to determine whether their use of the system or agent is negligent. [245] This concern

compared to a system that is not explainable. Thus, in general, developing or deploying an explainable system is inherently the more reasonable choice.

238. AI Act, *supra* note 117.

239. *See id.* art. 13 (addressing explainability); *id.* art. 14 (addressing human oversight).

240. *See also id.* art. 12 (imposing the EU's logging requirement, driven by the need to assess system compliance with the Act's provisions).

241. Selbst, *supra* note 3, at 1331.

242. A. Michael Froomkin et al., *When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning*, 61 ARIZ. L. REV. 33, 61 (2019).

243. *See* RESTATEMENT (THIRD) OF TORTS: PHYSICAL AND EMOTIONAL HARM § 7 (AM. L. INST. 2010).

244. Froomkin et al., *supra* note 242, at 61.

245. *See also* Selbst, *supra* note 3, at 1331–32. *Cf. id.* at 1360–61 (discussing situations where requirements of explainability or interpretability are not met); Padovan et al., *supra* note 54, at 133.

theoretically ties to the requirement that a breach of duty involves foreseeable harm.[246] In the context of AI systems and agents, these authors suggest that the unpredictability of the system makes it difficult to foresee the harm it might cause, making negligence liability seem less appropriate.[247] Similarly, when a human relies on an AI system or agent, the system's lack of explainability could render any resulting harm unforeseeable.[248]

Although foreseeability is often discussed under "causation,"[249] it is relevant to address it here. Whether the harm is foreseeable directly influences whether the deployer has acted with due diligence. Notably, AI systems or agents do not entirely undermine foreseeability. Authors who raise the "foreseeability problem" with current AI systems and agents do so because negligence liability depends on the negligent party's awareness of the potential for harm resulting from their actions.[250] However, negligence only requires that a category of harm be foreseeable, not the precise harm that occurs.[251]

---

246. Selbst, *supra* note 3, at 1332 (discussing the similar requirement under causation). *See generally* W. Jonathan Cardi, *Reconstructing Foreseeability*, 46 B.C. L. REV. 921 (2005) (discussing the foreseeability requirement).

247. *See, e.g.*, Selbst, *supra* note 3, at 1375; ANNA BECKERS & GUNTHER TEUBNER, THREE LIABILITY REGIMES FOR ARTIFICIAL INTELLIGENCE 73 (2021); *see also* Kowert, *supra* note 17, at 184 (descriptively, without adhering to that view).

248. *Cf.* Selbst, *supra* note 3, at 1331–32.

249. *Id.* at 1332.

250. *See id.* at 1346; *see also* David G. Owen, *Figuring Foreseeability*, 44 WAKE FOREST L. REV. 1277, 1286 (2009) (arguing that there should be such a potential appreciation); *cf.* Lemley & Casey, *supra* note 30, at 1311, 1315, 1378–79 (implicitly).

251. On the requirement of foreseeability of the category of harm, see Selbst, *supra* note 3, at 1342. *See generally* Cardi, *supra* note 246, at 925–26 (discussing the relevance of foreseeable injuries for the question of whether there was a breach of duty); Benjamin C. Zipursky, *Foreseeability in Breach, Duty, and Proximate Cause*, 44 WAKE FOREST L. REV. 1247, 1252 (2009) (discussing the Third Restatement).

For a discussion on how further limitations on the relevant behavior is imposed by the requirement of proximate cause, see Cardi, *supra* note 246, at 926. However, this limitation does not apply more strongly than that of the role of foreseeability for the breach of a duty, as this Article discusses it here.

Concerns about negligence liability for AI often seem to arise from technological exceptionalism rather than a fundamental issue.[252] AI's unpredictability does not undermine foreseeability, as both the possibility and category of harm remain clear. For instance, it is foreseeable that an AI agent may produce errors—a risk that developers should mitigate through warnings or deployer guidance. Likewise, the potential impact of such errors is predictable. Current AI systems are "weak" AI, limited to specific tasks that lead to readily identifiable, foreseeable harms—such as incorrect analyses or autonomous vehicle collisions.[253]

While greater explainability and foreseeability could aid in understanding the precise harm that might arise from reliance on an AI system's or agent's output, this level of specificity is not required to hold deployers liable. That said, this dynamic may shift[254] with the increasing autonomy of AI agents and potential future generations of "strong AI," which could exhibit entirely unforeseeable behaviors, potentially rendering the category of harm unpredictable.[255]

A precautionary focus on the negligent party's perspective reframes the issue of AI deployment. This approach recognizes that an AI deployer can foresee potentially harmful AI output when deciding to rely on it. As such, relying on AI output is comparable to engaging in risk in traditional contexts, where unreasonable risk-taking has long been recognized as a potential basis for negligence liability.[256]

---

252. *Cf. supra* note 81 and accompanying text (discussing causation).

253. Selbst, *supra* note 3, at 1343–44.
The distinction with strong AI is likely to blur as AI agents become capable of performing increasingly more distinct tasks, *see* Selbst, *supra* note 3, at 1344.

254. *Cf.* Ryan Calo, *Is the Law Ready for Driverless Cars?*, 61 COMM. ACM 34, 35 (2018). *See generally* Selbst, *supra* note 3, at 1343–43 (discussing the impact of increased AI autonomy).

255. *Cf.* Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 365 (2016) (not employing this terminology ("strong" AI)); Selbst, *supra* note 3, at 1344; Lai, *supra* note 4, at 629–30.

256. *See* RESTATEMENT (THIRD) OF TORTS: PHYSICAL AND EMOTIONAL HARM § 2 (AM. L. INST. 2010) (defining negligence using a risk-based approach); RESTATEMENT (SECOND) OF TORTS § 282 (AM. L. INST. 1965) (defining negligence using a risk-based approach); *cf.* Warren A. Seavey, *Principles of*

footnote continued on next page

Similarly, in medical malpractice, the standard does not require diligent physicians to save every patient.[257] Certainty is not a prerequisite for liability. The key issue is whether the risk engaged was impermissible. The Learned Hand[258] test provides an analogous framework for this evaluation, suggesting that reliance on an AI agent may be acceptable if additional precautions—such as further system training or some form of oversight—would impose costs disproportionate to the benefits expected from the system's or agent's improvement.[259] If these precautions are not unduly burdensome and are neglected, the developer's conduct (and risk engagement) should be deemed negligent. Moreover, the assumption of risk bars deployers from seeking liability against others, such as the system developer.[260]

The critical question is whether the deployer could reasonably rely on the specific AI agent. This involves several considerations. First, the deployer must select an AI agent appropriate for the intended function; choosing an agent clearly unsuitable for its purpose is generally likely to constitute negligence.[261] Second, the deployer must ensure diligent supervision of the AI agent during its use.[262]

Framing the challenge as an evaluation of reliance on the agent allows for a nuanced understanding of the varying degrees of deployer supervision. There is typically no issue when a deployer exercises

---

Torts, 56 HARV. L. REV. 72, 87 (1942); George P. Fletcher, *Fairness and Utility in Tort Theory*, 85 HARV. L. REV. 537, 557 n.74 (1972); Heidi Li Feldman, *Prudence, Benevolence, and Negligence: Virtue Ethics and Tort Law*, 74 CHI.-KENT L. REV. 1431, 1431, 1443 (2000).

257. *See* Froomkin et al., *supra* note 242, at 61.

258. United States v. Carroll Towing Co., 159 F.2d 169, 173 (2d Cir. 1947); *cf.* RESTATEMENT (THIRD) OF TORTS: PHYSICAL AND EMOTIONAL HARM § 3 (AM. L. INST. 2010). *See generally* White, *supra* note 216; Hylton, *supra* note 216; O'Gorman, *supra* note 216.

259. *Cf.* Vasant Dhar, *When to Trust Robots with Decisions, and When Not to*, HARV. BUS. REV. (May 17, 2016), https://hbr.org/2016/05/when-to-trust-robots-with-decisions-and-when-not-to [https://perma.cc/MH56-GE7N].

260. *See* Stein, *supra* note 47, at 986.

261. A reasonable person—the relevant standard here, *supra* note 228—would likely deploy an AI agent only if and if it is suited for the task at hand.

262. *Cf.* Ignacio N. Cofone, *Servers and Waiters: What Matters in the Law of A.I.*, 21 STAN. TECH. L. REV. 167, 191 (2018) (demonstrating that the author seems to assume that the supervisor should be the developer of the system).

diligence and uses an AI system as an additional safeguard, providing a separate analysis to cross-check some results.[263] In such cases, the deployer does not treat the system as a substitute for their own tasks. In fact, this approach can even be regarded as more diligent than instances where no AI system is used.[264]

The situation becomes more complex if the deployer relies entirely on the AI agent or system, rather than using it as a secondary check. Two critical factors must be considered: (1) the level of supervision exercised and (2) the depth of reliance on the system's "reasoning."

The first factor, supervision, is relatively straightforward. A deployer might thoroughly review every instance of AI output (as one may expect doctors to do when using a system to analyze vital scans). Alternatively, they might only verify notable results manually or, at the broadest level, review the agent's or system's output in general without scrutinizing any individual instances at all.

The second factor concerns the depth of reliance on the agent's analysis. For example, when a doctor reviews a patient's scan analyzed by an AI tool, they should apply their training and expertise to confirm the system's findings. If uncertain, the doctor should reflect on why the system may have reached a particular conclusion. This aspect closely relates to the explainability of the AI system and the deployer's level of expertise.[265] Especially for deployers without specific training or expertise—unlike medical professionals—the system's explanations should be sufficiently clear and simplified to be easily understood by a layperson. In such cases, these explanations may represent the only practical form of supervision for deployers lacking relevant expertise.[266]

---

263. Froomkin et al., *supra* note 242.

264. Using an AI system as an additional verification tool demonstrates due care and prudence, reducing the risk of errors and harm. In the same vein, the European Union has opted not to apply some of its requirements for high-risk AI systems in instances where the AI system is only used to improve the result of a previously completed human activity. *See* AI Act, *supra* note 13, art. 6(3)(b).

265. System explainability can aid the deployer in this analysis. Without adequate explainability, an independent analysis would be required. *Cf. supra* note 51.

266. *Supra* note 51 and accompanying text.

Increased supervision across both components—both deployers with and without relevant expertise—is generally a more diligent approach where feasible. However, this is not always practical. Autonomous vehicles highlight this limitation, as drivers cannot always override the system's actions.[267] Nevertheless, it remains prudent to deploy autonomous vehicles in a manner that allows for human intervention in certain circumstances, such as requiring a driver to remain in the driver's seat, as discussed earlier for developers.[268] More broadly, the decision to rely on the system or agent often takes a more abstract form, such as the initial choice to deploy the autonomous vehicle in the first place.[269]

Increasing supervision may be economically impractical for other AI agents, potentially undermining efficiency and cost advantages. While economic considerations heavily influence the extent of supervision,[270] the method of supervision depends significantly on the deployer's expertise and the system's explainability. If the deployer lacks deep expertise in the AI tool's specific domain, their ability to verify its outputs is inherently limited. In such situations—or when independent assessment is infeasible regardless of expertise—the deployer must rely on the system's or agent's justifications. As a result, the effectiveness of supervision is constrained by the system's explainability.

While the responsibility for ensuring explainability primarily lies with the AI developer, it is also a key consideration for the deployer during system selection. A deployer demonstrates greater diligence by choosing an AI agent with a higher degree of explainability, facilitating oversight wherever feasible.[271]

In general, increased supervision reduces reliance on an AI agent and alleviates concerns about negligence. This issue becomes

267. While many states require a human in the driver's seat, Kasap, *supra* note 233, a driver who has not been actively engaged for some time is unlikely to respond instantaneously to sudden, unexpected situations, *cf. infra* note 281 (discussing "alert fatigue"); Selbst, *supra* note 3, at 1347–48.

268. *Supra* Section IV.C.1.b.

269. *Cf.* Mbilike M. Mwafulirwa, *The Common Law and the Self-Driving Car*, 56 UNIV. SAN FRANCISCO L. REV. 395, 411 (2022).

270. *Cf. supra* note 218 and accompanying text.

271. Selbst, *supra* note 3, at 1331.

particularly acute with non-explainable AI agents, where it may be difficult or even impossible for the deployer to assess the system's accuracy in specific instances.[272]

For AI agents where oversight is not feasible—whether due to their inherent autonomy or a lack of explainability—the focus shifts significantly toward the decision to use the AI agent.[273] This extends beyond the *initial* decision to deploy the agent, also encompassing continued reliance after adoption.[274] When human oversight is exercised, reliance on the AI agent can become more specific and target particular outputs.[275] As discussed earlier, the deployer might use their expertise to evaluate the system's or agent's outputs. The broader question of reliance on the agent reemerges only when the deployer is unable or unwilling to perform such evaluations.

The potential lack of explainability is just one factor influencing whether reliance on the agent constitutes negligence. While it is generally more diligent to rely on an AI agent when its "reasoning" can be conceptually verified, this is not an absolute requirement, nor does it preclude the application of negligence liability.[276]

An intriguing scenario arises when an AI agent suggests or adopts a radically innovative approach that diverges from existing human practices and the deployer cannot independently determine its appropriateness.[277] In such cases, the critical question is not simply whether the deployer should have relied on the AI system, but rather, more holistically, whether the decision to adopt the system's

272. *Cf.* Selbst, *supra* note 3, at 1331–32.

273. In the same sense, see Selbst, *supra* note 3, at 1339; *cf.* Froomkin et al., *supra* note 242, at 61 (equating that decision to the decision to rely on a human).

274. *Cf.* Froomkin et al., *supra* note 242, at 61; Selbst, *supra* note 3, at 1339.

275. Depending on the extent and frequency of supervision exercised, the decision to rely on the agent's output may even concern a single specific output rather than the deployment of the agent as a whole. This can substantially influence the relevant risks and reasonableness of that reliance.

276. Explainability is inherently more difficult for some (complex) AI systems, *supra* Section II.B, and maybe fail to adequately offer insights, *supra* note 59, while those models may offer significant benefits.

277. *See, e.g.*, W. NICHOLSON PRICE II, *Medical Malpractice and Black-box Medicine*, *in* BIG DATA, HEALTH LAW, AND BIOETHICS 295, 300–01 (I. Glenn Cohen et al. eds., 2018).

unconventional suggestion was negligent under the given circumstances.[278]

There is a compelling argument for slightly relaxing the standard of care imposed on AI deployers in certain situations. This is particularly relevant for deployers overseeing AI systems during repetitive, routine tasks.[279] Humans are inherently ill-suited to maintain continuous attention and intervene as needed.[280] Similarly, when a human deployer must frequently override system warnings, they may become less vigilant in situations where the warnings should not be ignored.[281] This phenomenon, known as "alert fatigue," illustrates why it is not always desirable for actors to focus solely on limiting their own liability.[282] Excessive warnings from developers to deployers can further exacerbate this issue.[283] This Article's normative analysis in Part III also suggests that parties should not be incentivized to over-prioritize reducing their own liability at the expense of broader operational efficiency.[284]

Furthermore, negligence may not only serve as an argument against deploying AI but may, in some cases, necessitate its use. This is particularly evident in domains where, when deployed responsibly and with appropriate human oversight, AI systems surpass human capabilities.[285] It is also relevant when AI performance is comparable to human performance but offers significant efficiency gains. For instance, an AI agent that enables a service provider to deliver more affordable services exemplifies how the deployment of AI can be justified, provided it is conducted with diligence.[286]

---

278. For that assessment, the deployer should refer to their own expertise. *Cf. id.* at 301–03; *supra* note 51.

279. For a similar discussion, also discussing the possibility of a *higher* standard of diligence, see Selbst, *supra* note 3, at 1348–49.

280. In this sense, see *id.* at 1346; Smart et al., *supra* note 44, at 51.

281. Michael Greenberg & M. Susan Ridgely, *Clinical Decision Support and Malpractice Risk*, 306 JAMA 90, 90 (2011); Selbst, *supra* note 3, at 1347–48.

282. *See* Greenberg & Ridgely, *supra* note 281, at 90; Selbst, *supra* note 3, at 1347–48.

283. *Cf.* Greenberg & Ridgely, *supra* note 281, at 90; Selbst, *supra* note 3, at 1347–48.

284. *Supra* Part III.

285. Froomkin et al., *supra* note 242, at 61.

286. *Cf.* Cass v. 1410088 Ontario Inc., 2018 CanLII 6959, para 34 (Can. Ont. Sup. Ct. J.) (demonstrating how a lawyer failed to use AI tools to save time and, as a result, could not bill all the time spent to their client).

2. Products Liability

a. General

Products liability requires that a product be "defective" and that the defect was present when the product was placed under the buyer's control.[287] This presents unique challenges for AI,[288] as systems may continue training throughout their deployment.[289] This Article primarily focuses on cases where the deployer has not significantly contributed to the AI agent's training, though the discussion can similarly apply to other scenarios.

The primary challenge in applying products liability to AI agents lies in determining whether the system is defective. It is not enough for an erroneous system output to cause damage; the damage must result from a defect in the AI agent.[290] The inherent imperfection of AI agents and the possibility of occasional erroneous outputs do not necessarily imply defectiveness.

Products can be defective in various ways. Manufacturing defects are an example that can also occur in AI agents.[291] For example, a faulty sensor could constitute such a defect.[292] However, these issues are closely analogous to defects in traditional products and are less relevant to this Article's analysis. Instead, this discussion focuses on the more complex issue of the inherent imperfection of AI agents and systems.[293] Consequently, this Article does not address hardware defects, vulnerabilities that make agents susceptible to hacking, or the lack of safety in datasets.[294]

AI, however, blurs the traditional line between design and manufacturing defects. As previously noted, datasets containing

287. Lai, *supra* note 4, at 616.

288. *Id.*

289. When the agent is sold and placed under the buyer's control, any additional training by the deployer may complicate or even prevent proving that the defect existed at the time of sale.

290. Jean-Sébastien Borghetti, *How Can Artificial Intelligence Be Defective?*, *in* LIABILITY FOR ARTIFICIAL INTELLIGENCE AND THE INTERNET OF THINGS 63, 66 (S. Lohsse et al. eds., 2019).

291. Chagal-Feferkorn, *supra* note 4, at 86.

292. *Cf.* Funkhouser, *supra* note 93, at 453.

293. *Cf.* Griffin, *supra* note 39, at 92–93.

294. *See, e.g.*, Selbst, *supra* note 3, at 1350–51.

incorrect data can lead to erroneous AI outputs.[295] Whether datasets should be considered components of the AI system depends on how the "product" is defined.

One perspective would be to interpret the "AI system product" broadly, encompassing the parameters and operations the system uses to transform inputs to outputs, as well as the underlying model and training process that produced those parameters and operations. Alternatively, a narrower interpretation might define the "AI system product" as only the parameters and operations—the specific functions performed to generate output in a given instance. Such a restrictive view would acknowledge the dynamic nature of the "product," which evolves with each training instance and potentially during its use.

Under the broad interpretation, the "design" of the AI agent would include the algorithm, the training process, and the model used for training. In contrast, the narrower interpretation limits the design to the specific parameters and operations responsible for transforming input into output. From this perspective, training data becomes analogous to a component used in producing parameters and operations, making it similar to the manufacturing process. Consequently, flawed or substandard training data might be characterized as a manufacturing defect. Considering this, determining which interpretation of the AI "product" is most appropriate in a given case likely depends on the specific circumstances, such as whether the system was mainly trained by the developer or the deployer.

It is important to note, however, that incorrect training data is not the only source of erroneous AI output. Given the absence of established case law treating training data as a manufacturing defect—and considering the existence of other sources of incorrect AI outputs—it is worthwhile to explore the design defect regime in greater depth.[296]

295. *Supra* Section II.B.

296. *Cf.* Brian S. Haney, *The Optimal Agent: The Future of Autonomous Vehicles and Liability Theory*, 30 ALB. L.J. SCI. & TECH. 1, 28–31 (2020).

There are two primary tests used to determine whether a design defect exists:[297] the consumer expectations test[298] and the cost-benefit test,[299] the latter of which encompasses the "alternative design" test and the "failure to warn" doctrine outlined in the Third Restatement of Torts.[300] It is important to note that neither test requires absolute perfection in terms of safety.[301]

Some courts apply one of these tests while others consider both.[302] In the latter case, the availability of an alternative design can be a factor in assessing whether a product is unreasonably dangerous under the consumer expectations test.[303] Similarly, some commentators emphasize the overlap between these two approaches.[304] For clarity, this Article discusses each approach in turn.

b. Defects Under the Consumer Expectations Test

The consumer expectations test is perhaps the most straightforward. It evaluates whether the product performs as safely and effectively as an ordinary consumer would expect.[305] A product is deemed defective if it falls short of this standard.[306] However, determining what an ordinary consumer might expect from an

297. Selbst, *supra* note 3, at 1323; LEWIS BASS & THOMAS PARKER REDICK, PRODUCTS LIABILITY: DESIGN & MANUFACTURING DEFECTS § 4:1 (2d ed. 2024); SPEISER ET AL., *supra* note 205, at §§ 18:5–6.

298. SPEISER ET AL., *supra* note 205, at § 18:6.

299. *Id.* at § 18:5.

300. RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 2 (AM. L. INST. 1998); Aaron D. Twerski & James A. Henderson, Jr., *Manufacturers' Liability for Defective Product Designs: The Triumph of Risk-Utility*, 74 BROOK. L. REV. 1061, 1064–65 (2009); Emily Frascaroli et al., *Let's Be Reasonable: The Consumer Expectations Test is Simply Not Viable to Determine Design Defect for Complex Autonomous Vehicle Technology*, 2019 J. L. & MOBILITY 53, 59 (2019) (discussing the alternative design test).

301. BASS & REDICK, *supra* note 297, at § 4:1.

302. BASS & REDICK, *supra* note 297, at § 4:12.

303. McCourt v. J.C. Penney Co., Inc., 734 P.2d 696, 698 (Nev. 1987).

304. Selbst, *supra* note 3, at 1324.

305. Vladeck, *supra* note 3, at 134; BASS & REDICK, *supra* note 297, at § 4:1; SPEISER ET AL., *supra* note 205, at § 18:6. *Cf.* Selbst, *supra* note 3, at 1323–24.

306. Vladeck, *supra* note 3, at 134; BASS & REDICK, *supra* note 297, at § 4:1; SPEISER ET AL., *supra* note 205, at § 18:6. *Cf.* Selbst, *supra* note 3, at 1323–24.

inherently imperfect AI agent is challenging. These expectations must be *reasonable*, as the test assumes an ordinary, reasonable consumer.[307]

This raises the question of whether manufacturers can shield themselves from liability by providing advance warnings to consumers. This issue is more directly addressed in the cost-benefit test.

c. Defects Under the Cost-Benefit Test

The cost-benefit test, also known as the "risk-utility test,"[308] is closely aligned with a negligence analysis, employing principles similar to the Learned Hand formula.[309] The Third Restatement of Torts adopts this approach in § 2:

> A product is defective . . . in design when the foreseeable risks of harm posed by the product could have been reduced or avoided by the adoption of a reasonable alternative design by the seller or other distributor . . . and the omission of the alternative design renders the product not reasonably safe.[310]

Under this test, a product defect is found when the developer could have reasonably reduced or avoided foreseeable risks by adopting an alternative design, rendering the product "unreasonably unsafe."[311] Notably, the usefulness and desirability of the product—two essential policy considerations—are also taken into account when determining whether it is defective.[312]

The scope of "alternative designs" is not inherently limited and may include non-AI products.[313] However, this is where the added value of the AI product becomes crucial: If a product lacks the unique

307. *Cf.* Selbst, *supra* note 3, at 1324 (implicitly, for autonomous vehicles).
308. Vladeck, *supra* note 3, at 135.
309. Scott, *supra* note 177, at 467; Twerski & Henderson, *supra* note 297, at 1064–65; Chagal-Feferkorn, *supra* note 4, at 81.
310. RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 2 (AM. L. INST. 1998).
311. *Cf.* Griffin, *supra* note 39, at 79.
312. Wade, *supra* note 174, at 837.
313. Griffin, *supra* note 39, at 85.

advantages an AI tool provides, it may not qualify as a reasonable alternative.[314]

Reasonable alternative designs may also include AI agents or systems employing different techniques, training data, or approaches.[315] While these alternatives may mitigate risk, they do not necessarily eliminate it entirely. They can, however, meaningfully reduce it. More feasible alternatives might also involve designs incorporating measures to enhance human oversight, such as improved user interfaces.[316]

The concept of reasonableness sets an important boundary. As discussed earlier in Section II.B, requiring continuous investment in an AI agent for marginal performance improvements imposes an unreasonable burden beyond a certain point.[317] Such investments should only be pursued to the extent they remain reasonable. A negligence-based analysis supports this limitation, with the Learned Hand test offering valuable guidance.[318]

Determining whether an alternative design is "reasonable" illustrates the challenges of applying this test to complex AI agents. For example, in the case of an autonomous vehicle crash, a victim would need to show that it was reasonable for the developer to anticipate and test for the specific crash scenario.[319] While some argue that the cost of additional testing is minimal compared to potential damages,[320] the unpredictability of specific crash scenarios—and the difficulty of determining which scenarios to test for—may make comprehensive testing unattainable or prohibitively expensive.[321] This

314. *Cf. id.* at 87 ("[F]or many AI systems, the argument that humans alone are better as a [reasonable alternative design] will likely fail because even when AI systems have notable issues, they still often outperform humans alone.").

315. *Id.*

316. *Id.* (treating this as a distinct category).

317. *Supra* Section II.B.

318. Griffin, *supra* note 39, at 88.

319. Selbst, *supra* note 3, at 1324.

320. *Cf.* Marchant & Lindor, *supra* note 175, at 1334. *See generally* F. Patrick Hubbard, "*Sophisticated Robots*": *Balancing Liability, Regulation, and Innovation*, 66 FLA. L. REV. 1803, 1854–55 (2014) (discussing the role of expert testimony); Selbst, *supra* note 3, at 1325 (discussing how many scenarios would have to be involved in some tests).

321. *Cf.* Hubbard, *supra* note 320, at 1854–55; Selbst, *supra* note 3, at 1325.

complexity may render it "unreasonable" to expect developers to identify and test for every potential improvement to their AI models.

A safer, reasonable alternative design, though necessary, is not sufficient for a product to be deemed defective.[322] An AI agent is considered defective only if the increased risk from its original design, compared to the reasonable alternative, renders it "not reasonably safe."[323] This standard is critical yet inherently ambiguous, as AI agents, like many products, naturally involve some foreseeable risks.[324]

The consumer expectations test may be relevant here, as the expectations of an ordinary consumer are sometimes used to assess the reasonableness of a product's safety.[325] Another possible criterion for evaluating reasonableness is to compare the risks posed by the AI agent to those involved in similar activities conducted without AI agents, although this should not be the sole benchmark.[326]

The flexibility of this requirement highlights its alignment with the negligence regime. It also underscores the challenge of applying the cost-benefit test to complex AI agents, as victims may struggle to prove an AI agent is "not reasonably safe" when confronted with highly sophisticated and opaque technology.

d. Warning and Defectiveness

In addition to design defects, a product may also be deemed defective if it lacks adequate warnings of foreseeable risks, rendering it not reasonably safe.[327] According to § 2(c) of the Third Restatement of Torts,

> [a product] is defective because of inadequate instructions or warnings when the foreseeable risks of harm posed by the product could have been reduced or

322. Selbst, *supra* note 3, at 1323.

323. *Cf.* Griffin, *supra* note 39, at 88.

324. *Cf. id.* at 79, although this Article is not limited to the list of foreseeable risks provided there.

325. *Id.* at 88.

326. Some authors stress that the test should differ from a comparison to human behavior. Vladeck, *supra* note 3, at 132; Griffin, *supra* note 39, at 89.

327. RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 2 (AM. L. INST. 1998); RESTATEMENT (SECOND) OF TORTS § 402A cmt. k (AM. L. INST. 1965); *cf.* Griffin, *supra* note 39, at 93.

> avoided by the provision of reasonable instructions or warnings by the seller or other distributor . . . and the omission of the instructions or warnings renders the product not reasonably safe.[328]

While this type of defect is closely related to design defects—sharing the familiar standard of "not reasonably safe"—it is often treated as a distinct category of defect,[329] as acknowledged by the Third Restatement.[330]

This distinction suggests that a developer of an AI agent perceived as unsafe may avoid products liability by providing adequate warnings about the agent's risks.[331] For such warnings to be effective, they must be directed to the individual using the agent or system and also to the parties responsible for overseeing its deployment,[332] rather than to the person harmed by the product.[333] In medical contexts, for instance, it may suffice for the manufacturer to warn the physician.[334] This approach recognizes the difficulty manufacturers often face in directly reaching the end user;[335] however, in general, the end user should still receive adequate warnings.[336]

---

328. RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 2(c) (AM. L. INST. 1998).

329. Selbst, *supra* note 3, at 1323.

330. RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 2(a), (c) (AM. L. INST. 1998).

331. *See generally* RESTATEMENT (SECOND) OF TORTS § 402A, cmt. k (AM. L. INST. 1965) ("Such a product, properly prepared, and accompanied by proper directions and warning, is not defective, nor is it *unreasonably* dangerous."); Taylor v. Intuitive Surgical, Inc., 389 P.3d 517 (Wash. 2017); Patrick H. O'Neill, Jr., *Unavoidably Unsafe Products and the Design Defect Theory: An Analysis of Applying Comment K to Strict Liability and Negligence Claims*, 15 WM. MITCHELL L. REV. 1049, 1055 (1989); Roe, *supra* note 10, at 340; Griffin, *supra* note 39, at 93–94.

332. *Taylor*, 389 P.3d at 522 (indicating that the warning must be directed at the buyer, based on Section 7.72.030 (1) of the Washington Product Liability Act).

333. Terhune v. A.H. Robins Co., 90 Wash. 2d 9, 13 (1978).

334. *Id.* at 13.

335. *Id.*

336. *Taylor*, 389 P.3d. at 525–26.

e. Strict Liability?

It is important to note that the design defect regime does not generally impose strict liability.[337] Under the cost-benefit analysis, determining a defect is primarily a test of reasonableness.[338] If a product is not reasonably safe, its development is deemed negligent if a reasonable alternative design exists but was not adopted.[339] The consumer expectations test, by contrast, is somewhat stricter. It does not require negligence in the product's design but instead imposes the standard based on what a reasonable consumer would find adequate.[340] Developers can, however, shape consumer perceptions by providing sufficient information and warnings about the product.[341]

As a result, the design defect regime is not inherently stringent. This represents a shift from the historical approach to products liability. Under the Second Restatement of Torts, there was no requirement to demonstrate the availability of a reasonably safe alternative design.[342] This stricter traditional approach continues to apply to manufacturing defects: A product is considered to have a manufacturing defect when "the product departs from its intended design even though all possible care was exercised in the preparation and marketing of the product."[343] By explicitly stating that liability applies regardless of the level of care, this regime is more aligned with strict liability principles. This underscores the potential implications of classifying flawed data as a manufacturing defect.

---

337. Some authors have recently discussed products liability for AI systems as if it necessarily entails a strict liability regime. However, their arguments primarily rely on the outdated *Restatement (Second) of Torts*. *See, e.g.*, Lai, *supra* note 4, at 615.

338. As implied by the terminology "*reasonable* alternative design" in RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 2(b) (AM. L. INST. 1998) (emphasis added).

339. *Id.*

340. Vladeck, *supra* note 3, at 134; BASS & REDICK, *supra* note 297, at § 4:1; SPEISER ET AL., *supra* note 205, at § 18:6.

341. *Cf.* BASS & REDICK, *supra* note 297, at § 4:12 (implied).

342. *See* RESTATEMENT (SECOND) OF TORTS § 402A (AM. L. INST. 1965).

343. RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 2 (AM. L. INST. 1998).

D. Causation

Causation[344] poses a significant challenge in applying tort law to harm caused by AI agents.[345] This difficulty is particularly pronounced in cases of negligence by AI developers but also extends to deployers.[346] For example, consider a developer who creates an AI agent that fails to meet industry standards, resulting in harm caused by erroneous outputs. How can the victim prove that the harm would not have occurred but for the developer's inadequate programming or training? Even a well-designed and properly trained agent or system can occasionally produce erroneous outputs that lead to harm, making causation difficult to establish.[347]

For deployers, the breach of duty typically involves relying on the agent when such reliance is unwarranted. The alternative scenario would be for the deployer to have avoided relying on the agent. However, this does not necessarily mean the harm would not have occurred; humans, too, are prone to error. Therefore, the mere occurrence of harm does not automatically establish negligence.[348]

Causation is equally critical in products liability,[349] which requires that the defective product caused the harm.[350] However, strict products liability regimes provide a distinct advantage, as they do not require proof that the developer's conduct is what caused the harm.[351]

---

344. While this discussion focuses on the causation requirement, similar challenges arise from the foreseeability requirement, which also indirectly limits the scope of causation.

345. Lai, *supra* note 4, at 628; Zhao Yan Lee et al., *Deep Learning Artificial Intelligence and the Law of Causation: Application, Challenges and Solutions*, 30 INFO. & COMMC'NS TECH. L. 255, 265 (2021).

346. *See, e.g.*, Lai, *supra* note 4, at 613. On the challenging nature of causality as to system deployers, *cf.* Gerstner, *supra* note 3 (noting how the complexity of AI systems and multiple parties complicate the analysis), at 249; *see* Lai, *supra* note 4, at 628.

347. *See supra* Section II.B on the inherent imperfection of AI.

348. *See, e.g.*, Froomkin et al., *supra* note 242, at 61 (discussing this for medical professionals).

349. Abbott, *supra* note 96, at 13.

350. Roe, *supra* note 10, at 331–32; BASS & REDICK, *supra* note 297, at § 4:40.

351. Instead, they only require that the defect caused the harm, regardless of whether the developer could have prevented it. In non-strict products

footnote continued on next page

This circumvents the challenges of evaluating the impact of more diligent development or assessing the foreseeability of the harm from a causation perspective.

When damage occurs following some intervention or use by the deployer, a recurring issue in products liability is determining whether the agent caused the harm.[352] This is particularly complex in cases where human oversight influences the agent's or system's functioning.[353]

Causation is often seen as a major challenge in negligence liability, with some arguing that existing frameworks are ineffective for certain AI systems.[354] This concern, largely tied to AI's "black-box" nature,[355] should not be overstated. While the lack of explainability does introduce specific challenges,[356] these are addressed mainly by the preceding discussion on foreseeability. As noted, tort law does not require the specific harm to be foreseeable.[357] Rather, it is sufficient that the *category* of harm is foreseeable, which is generally the case for contemporary AI systems and agents.[358]

The challenges of causation must be examined with some nuance. To begin, causation need not be proven with absolute certainty. It suffices to establish that it is "more likely than not."[359] This standard, significantly lower than in some legal systems that require a threshold

liability regimes, causation plays a more nuanced role, rather resembling its function in negligence liability.

352. *Cf.* Roe, *supra* note 10, at 331–32.

353. *Id.*

354. *See, e.g.*, Zhao Yan Lee et al., *supra* note 345, at 262 (discussing deep-learning AI).

355. Some authors emphasize this link. *See, e.g.*, Lai, *supra* note 4, at 628–29; Sylwia Wojtczak & Paweł Księżak, *Causation in Civil Law and the Problems of Transparency in AI*, 29 EUR. REV. PRIV. L. 561, 581 (2021).

356. *See, e.g.*, Lai, *supra* note 4, at 628–29 (2021).

357. *Supra* Section IV.C.

358. *Supra* Section IV.C.1.c.

359. In the context of medical damage (potential development of cancer), see Gideon v. Johns-Manville Sales Corp., 761 F.2d 1129, 1138 (5th Cir. 1985); Robert T. Ebert, Jr., *Damages for an Increased Risk of Developing Cancer Caused by Asbestos Exposure Are Only Recoverable If It Is More Likely Than Not That Cancer Will Develop*, 51 MO. L. REV. 847, 848 (1986).

approaching ninety percent,[360] substantially reduces the burden of proof.

In addition, this issue is not unique to AI agents. In medical cases, for instance, it is often unclear how a practitioner's breach of duty affects a patient's recovery. Two key approaches address this causality challenge. First, some argue that the "loss of chance" doctrine could apply in cases of uncertain causation,[361] allowing compensation based on a reduced probability of a desirable outcome or an increased risk of harm.[362] In some contexts, this theory has been used to lower the evidentiary threshold for proving causation.[363] More broadly, "probabilistic causality"[364] permits compensation despite uncertain causation.[365] While both could help victims of AI-induced harm recover damages from deployers or developers, these theories are mainly applied in medical malpractice cases.[366]

In fields like medical malpractice, these doctrines should logically be extended to cover the deployment of AI systems and agents. However, *all* victims of AI-related harm will likely face similar general challenges in proving causation. As a result, victims outside of the medical malpractice context will struggle to receive compensation.

---

360. *See* Mark Schweizer, *The Civil Standard of Proof—What Is It, Actually?*, 20 INT'L J. EVIDENCE & PROOF 217, 220 (2016).

361. *See, e.g.*, Dillon v. Evanston Hosp., 199 Ill. 2d 483, 504 (2002) (discussing the possibility of incurring future injuries due to negligent behavior); Robert S. Bruer, *Loss of a Chance As a Cause of Action in Medical Malpractice Cases*, 59 MO. L. REV. 969, 973 (1994); Timothy Dylan Reeves, *Tort Liability for Manufacturers of Violent Video Games: A Situational Discussion of the Causation Calamity*, 60 ALA. L. REV. 519, 543–44 (2009); Robert J. Rhee, *Loss of Chance, Probabilistic Cause, and Damage Calculations: The Error in Matsuyama v. Birnbaum and the Majority Rule of Damages in Many Jurisdictions More Generally*, 1 SUFFOLK U. L. REV. ONLINE 39, 39 (2013).

362. David A. Fischer, *Tort Recovery for Loss of a Chance*, 36 WAKE FOREST L. REV. 605, 606 (2001); Rhee, *supra* note 361, at 41–42.

363. RESTATEMENT (SECOND) OF TORTS § 323(a) (AM. L. INST. 1965); Bruer, *supra* note 361, at 974–75.

364. *See* Glen O. Robinson, *Probabilistic Causation and Compensation for Tortious Risk*, 14 J. LEGAL STUD. 779, 780–81 (1985); Alessandro Romano, *God's Dice: The Law in a Probabilistic World*, 41 U. DAYTON L. REV. 57, 75 (2016).

365. Romano, *supra* note 364, at 75.

366. Fischer, *supra* note 362, at 605–06; Smith & Fotheringham, *supra* note 93, at 143–44.

One solution would be to extend the medical malpractice regime. Another solution, formerly proposed by the European Union, would be to shift the burden of proof.[367] At a minimum, such a shift would incentivize developers and deployers to maintain meticulous system performance logs, allowing them to better mitigate potential liability claims. The European Union's AI Act explicitly includes such logging requirements.[368]

Alternatively, a similar outcome could be achieved by explicitly incorporating logging as part of the duty to develop or deploy AI agents and systems diligently. Arguably, this requirement already exists to some extent: If developers and deployers fail to monitor the AI agent's performance and outputs, they may find it difficult to demonstrate their diligence when defending against a claim, particularly if some elements suggest that they may have acted negligently.

A related issue is whether the deployer's negligence can absolve the developer from liability. Tesla notably invoked this defense in a case involving an accident with a semi-autonomous vehicle.[369] Under the sometimes-criticized[370] theory of "enabling torts," negligence that facilitates the negligence of others can also constitute a tort.[371] Consequently, it is unlikely that the concept of *novus actus interveniens*[372] will play a significant role in AI liability law.[373] Furthermore, it would be undesirable from a societal perspective to

367. *See Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive)*, art. 4, COM (2022) 496 final (Sept. 28, 2022).

368. *See* AI Act, *supra* note 13, art. 12.

369. Kowert, *supra* note 17, at 184.

370. *See* John C. P. Goldberg & Benjamin C. Zipursky, *Intervening Wrongdoing in Tort: The Restatement (Third)'s Unfortunate Embrace of Negligent Enabling*, 44 WAKE FOREST L. REV. 1211, 1218–31 (2009).

371. RESTATEMENT (THIRD) OF TORTS § 34 (AM. L. INST. 2012); Robert L. Rabin, *Enabling Torts*, 49 DEPAUL L. REV. 435, 436–453 (2000).

372. Under English law, gross negligence of a subsequent actor absolves the first actor of their liability for negligence. *See* Smith & Fotheringham, *supra* note 93, at 145.

373. Kowert, *supra* note 17, at 184.

incentivize developers to design agents and systems that emphasize deployer negligence rather than minimize the potential for harm.[374]

V. DECIPHERING THE AI LIABILITY DEFICIT

The preceding doctrinal discussion enables this Article to evaluate how well this regime aligns with the preferences outlined in the normative analysis. To structure this assessment, it is useful to differentiate between AI developers and AI deployers. While these categories do not fully reflect the complexity of the AI supply chain, they offer a clear framework for understanding the liability regime applicable to many key actors.

A. Liability Standards

1. Developers

A normative assessment suggests that developers should bear strict liability for all harm caused by their AI agents. However, this should not be the case when the harm results from a deployer's informed and voluntary decision or when significant benefits are externalized. Both exceptions assume the victim has not acted negligently.

Imposing strict liability on developers would incentivize them to provide adequate warnings and information to deployers and improve their AI agents' performance. Moreover, it would encourage developers to prioritize features such as explainability, reducing the expected harm even when deployers exercise reasonable care.

However, the current products liability regime falls short of these objectives. First, it is limited to addressing non-economic harm.[375] Second, it can feel outdated, as seen in the historical uncertainty over whether AI systems and agents qualify as products.[376] This issue is particularly evident in the distinction between manufacturing and design defects, where its application to AI systems appears somewhat forced.[377] While the duty to warn is increasingly recognized (especially

374. *Cf.* Smith & Fotheringham, *supra* note 93, at 143–44. *See generally* Bruer, *supra* note 361, at 971–72 (1994).

375. *Supra* Section IV.A.

376. *Supra* Section IV.B.

377. *Supra* Section IV.C.2.a.

for design defects),[378] erroneous or undesirable AI outputs due to the inherent imperfection of AI systems are not classified as manufacturing defects.[379] As a result, strict liability does not apply. Instead, such outputs are typically treated as design defects, subject to a reasonableness standard rather than strict liability.[380] This often shifts the cost of AI-caused harm onto the victim or deployer, leading to economic inefficiency.[381]

Furthermore, the products liability regime fails to adequately incentivize developers to optimize AI performance or incorporate critical features like explainability. While this may partially offset the externalization of benefits associated with certain AI agents and systems, this "compensation" is overinclusive, encompassing situations where no such externalization occurs.[382] Moreover, in cases where no liability attaches, for example, due to the reasonable behavior of the system developer, the victim bears the full cost of the harm. Normative analysis suggests a more desirable approach would distribute this harm across society rather than leaving the victim as the sole bearer.[383]

While AI developers are also subject to negligence liability, this regime is not inherently designed to establish strict liability. Instead, it reinforces certain outcomes under products liability, given the significant overlap between both regimes in assessing the reasonableness of design defects.[384] Negligence liability also underscores the need to provide adequate warnings to AI deployers.[385] As a result, both regimes fall short of establishing the desirable strict liability framework even when taken together.

This highlights the need to better align the current developer regime with principles of economic efficiency and fairness, ensuring burdens and incentives are properly distributed across the AI supply

378. *Supra* Section IV.C.2.d.
379. *Supra* Section IV.C.2.a.
380. *Supra* Sections IV.C.2.a–IV.C.2.e.
381. *Supra* Section III.C.
382. That externalization is evident in various AI systems that offer significant societal benefits beyond those directly involved. Examples include autonomous vehicles and AI-driven drug development. *Supra* Section III.B.
383. *Supra* Section III.C.
384. *Supra* Section IV.C.2.c.
385. *Supra* Section IV.C.1.b.

chain. One approach is to classify poor training data—or, more broadly, inadequate training—as a manufacturing defect, which could help address key shortcomings in the existing regime. In many cases, this shift would impose strict liability, fostering a fairer and more effective legal framework.

Alternatively, the existing regime could be more closely aligned with strict liability through a rigorous application of the open-ended "reasonable" alternative design test.[386] From an economic perspective, a stringent cost-benefit analysis under this test is particularly crucial when a traditional human alternative exists. In such cases, the "alternative design" should also consider excluding AI entirely.

These adjustments underscore the need for regulatory intervention in cases where AI systems and agents provide societal benefits that extend beyond their direct deployers, as seen with autonomous vehicles.[387] Under the existing regime—and even more so if stricter rules are imposed on developers—the associated risks and costs fall on the AI supply and deployment chain, while the benefits are externalized.[388] In such cases, a policy correction is needed to distribute the harms caused by these AI systems more equitably across the population that benefits from them, thereby preventing undesirable incentives for developers and deployers—such as through a mandatory insurance scheme.[389]

2. Deployers

The situation is more straightforward for AI agent deployers, who are subject only to negligence liability.[390] However, a significant challenge lies in the absence of an overarching duty of care for deployers. Instead, the analysis is highly context-specific, depending on whether it involves an autonomous vehicle, a medical tool used by a doctor, or an AI chatbot providing legal advice. Nonetheless, the preceding analysis allows for general observations to guide more specific evaluations of a given duty of care.

---

386. *Supra* Section IV.C.2.c.

387. *Supra* Section III.B. For a discussion of the lack of inherent corrective mechanisms in the general liability regime, see *supra* Part IV.

388. *Supra* Sections III.B and IV.

389. *Supra* Sections III.C and IV.

390. *Supra* Section IV.C.1.

A key insight from the normative assessment is that the intensity of the duty of care should depend on how free and informed the deployer was in their choice and use of the AI agent. For example, suppose the impact of an AI agent or system is relatively straightforward—such as a classification tool used by a doctor—and the deployer has been thoroughly informed about the system's limitations. In such cases, one can assume they made a free and informed choice; the doctor should bear ultimate liability if the system or agent output proves harmful. Therefore, the reasonableness of reliance by well-warned and well-informed deployers should be evaluated more strictly, approaching a form of strict liability. This is particularly applicable when the deployer possesses relevant experience or expertise, enabling them to better assess the risks associated with the AI agent's use.

However, many AI agents and systems are too complex for such an approach. Autonomous vehicles provide a clear example: They involve numerous potential issues across a vast array of scenarios, making it nearly impossible to make a fully informed decision about relying on the agent.[391] In such cases, the required standard of care is less stringent and varies according to circumstances. The more informed a deployer is, the greater their responsibility and the higher the standard of care required.[392]

How informed a deployer is depends not only on the information provided by the developer but also on factors such as the AI agent's inherent complexity, its deployment context, and the deployer's experience and expertise in the relevant domain.

The existing negligence liability regime addresses some of these factors. As deployers make more informed decisions about using an AI agent, their responsibility naturally increases.[393] This aligns with the principle of risk engagement: The more informed a deployer is, the

391. *Supra* Section III.A.

392. This principle applies more broadly, including to situations involving involuntary victims of an AI agent. While this Articles does not explicitly analyze such cases here, when those victims are aware of a particular risk posed by the system, the key question becomes how they manage that risk. *See, e.g.*, Cole, *supra* note 194, at 226–27.

393. *Supra* Section IV.C.1.b.

greater their duty to take precautionary measures such as supervising the agent or system.[394] The Learned Hand formula suggests an upper limit on these obligations, indicating that deployers are not required to take precautions when their costs outweighs the expected harm.[395] Furthermore, deployers' knowledge and understanding of the risks may indicate that they have been adequately warned by the system developer, potentially weakening a liability claim against them.

However, by capping the standard of care required of deployers, the current regime can undermine the goal of imposing greater responsibility on them. This limitation is only partially mitigated in scenarios where oversight is prohibitively expensive or impractical, such as in some cases involving autonomous vehicles. These high-complexity situations make it difficult to determine whether the deployer made a truly well-informed decision. As a result, the current regime may fail to provide adequate incentives for deployers to act responsibly.

B. Causation

Finally, AI agents pose general challenges regarding the causation requirement, particularly within negligence liability. While these challenges do not necessarily preclude successful liability claims—as seen with medical AI tools where probabilistic causality is more widely accepted—they do highlight the potential benefits of expanding the scope of probabilistic causality regimes. Alternatively, legislative interventions, such as the shifted burden of proof formerly proposed in the European Union, could presumptively equate the behavior of the agent or system with that of the developer, simplifying causation issues.[396]

VI. Conclusion

AI undoubtedly brings legal complexities, but it also offers significant advantages, including efficiency gains,[397] cost savings,[398]

394. *Supra* Section IV.C.1.c.
395. *Supra* Section IV.C.1.c.
396. *See* Proposal for an AI Liability Directive, *supra* note 367, art. 4.
397. *Supra* Section II.B.1.
398. *Supra* Section II.B.1.

and, in many cases, superior outcomes.[399] While concerns about unpredictable results,[400] opacity,[401] and occasional extreme errors merit attention, these properties of contemporary AI agents do not fundamentally upend tort law.[402] Rather, economic analysis suggests that traditional principles[403] can adapt to AI's unique challenges.

Key adjustments include recognizing that AI's externalized benefits may warrant legislative tools, such as insurance schemes[404] or subsidies,[405] and that the current products liability framework—which hinges upon a distinction between design and manufacturing defects—does not always map neatly onto AI and its reliance on data.[406] In some cases, stricter liability rules might incentivize appropriate care better than a traditional negligence-based approach.

Despite these challenges, existing negligence principles already encourage developers to ensure accuracy, explainability, and effective oversight. Provisions in the European Union's AI Act reflect these obligations and demonstrate that a nuanced reading of current law can often address AI's emerging issues without necessitating an entirely new regime. Ultimately, careful refinement, rather than wholesale replacement, of existing legal frameworks—coupled with ongoing technological literacy among legal professionals—will help the law adapt to AI's rapid evolution.

399. *Supra* Section II.B.1.
400. *Supra* Section II.B.2.
401. *Supra* Section II.B.2.
402. *Supra* Section III.C.
403. More particularly, the economic analysis underpinning tort law for traditional tools and products largely applies to AI systems and agents as well, *supra* Section III.C.
404. *Supra* Section III.C.
405. *Supra* Section III.C.
406. *Supra* Section IV.C.2.a.