Tort Law at the Frontier of Artificial Intelligence

Ketan Ramakrishnan[†]

The frontier of contemporary AI development is dominated by AI systems built on foundation models—highly versatile algorithms, trained in the first instance on broad swaths of data, that can function as speakers, tools, and agents across a wide variety of commercial, social, military, and political domains. For the moment, at least, the process of developing and releasing foundation models is subject to anemic ex ante regulation. Until that changes, it is largely the common law of torts—our society's most general legal mechanism for governing serious risks of physical injury—that will govern the frontier of AI development.

This Article offers an in-depth conceptual, doctrinal, and normative examination of tort liability for foundation-model development and release. It provides a qualified defense of the tort of negligence, the common law's most general and flexible cause of action, as the principal doctrinal foundation of the tort system's governance of this novel domain. Legal scholarship on AI liability—much of which predates the advent of foundation models—has been largely hostile to negligence, arguing that AI should principally or exclusively be governed by alternative doctrinal regimes. By contrast, this Article argues that the generality and flexibility of the negligence tort—and its greater sensitivity to the externalized benefits of risky activity—render it better suited to the polymathic and protean character of foundation models and the complex technical and institutional dynamics of their development and release.

---

† Associate Professor of Law, Yale Law School. For valuable suggestions and conversations, I am grateful to Elina Alter, Ricky Altieri, Leopold Aschenbrenner, Ian Ayres, Shyam Balganesh, Jack Balkin, Dean Ball, Avital Balwit, Nick Beckstead, Adam Billen, Mitch Berman, Nathan Calvin, Joe Carlsmith, Mala Chatterjee, Paul Christiano, Beba Cibralic, Ajeya Cotra, Tino Cuellar, Ben Eidelson, William Eskridge, Kim Ferzan, Koji Flynn-Do, Owen Gallogly, Mark Geistfeld, Sunny Gandhi, Abbe Gluck, John Goldberg, Jack Goldsmith, Jeff Gordon, Katja Grace, Oona Hathaway, Dan Hendrycks, Tim Hwang, Ari Kagan, Becca Kagan, Holden Karnofsky, Cullen O'Keefe, Saif Khan, Arden Koehler, Daniel Kokotajlo, Seb Krier, Doug Kysar, Nikita Lalwani, Larry Lessig, Gideon Lewis-Kraus, Josh Macey, Daniel Markovits, Martha Minow, John Morley, Ben Murphy, Chris Painter, Robert Post, Claire Priest, Vivek Ramakrishnan, Cristina Rodriguez, Roberta Romano, Steve Sachs, Peter Salib, David Schleicher, Tim Schnabel, Scott Shapiro, Yo Shavit, Divya Siddharth, Sandy Steel, Helen Toner, Alex Traub, Matt Wansley, Gabe Weil, Garrett West, Rebecca Wexler, John Witt, Diane Wood, Thomas Woodside, Felix Wu, Gideon Yaffe, Christopher Yoo, Taisu Zhang, Ben Zipursky, Jonathan Zittrain, and participants in faculty workshops at Cardozo and Yale. For excellent research assistance and invaluable editorial support, I am deeply indebted to Jared Riggs, Ivana Bozic, and Chinmay Deshpande, as well as Nashely Alvarez, Jackson Dellinger, Ara Johnson, and the other editors of the *Yale Journal on Regulation*.

Analyzing the choice between negligence and competing doctrinal regimes does, however, reveal important ways in which the courts should incrementally develop the law of negligence in order to properly address the nature and risks of foundation models. First, courts should expand the scope of the duty of care in negligence in order to provide redress when foundation models cause economic or emotional injury by behaving in a manner closely analogous to serious human wrongdoing. Second, courts should import into the law of negligence a malfunction doctrine, one similar to that which already exists in products liability. When a model behaves in a manner that is closely analogous to an intentional tort or crime, or that clearly contravenes legitimate expectations of safe performance, courts should allow the plaintiff to reach the jury even without adducing (further) evidence on breach.

The Article's analysis also suggests certain limitations and pathologies of tort liability as a mechanism of AI governance—issues that cannot be adequately addressed by any amount of judicial doctrinal development or legislative modification of the applicable liability rules. In particular, the specter of tort liability can be expected to disincentivize frontier AI developers from investigating and disclosing the novel and poorly understood risks that frontier-AI development may pose. That is especially disturbing given that our society is—unavoidably, to a large extent—relying quite heavily upon frontier AI developers themselves in order to discover, understand, and mitigate these risks. So, as a mechanism of frontier-AI governance, ex post liability alone is not only inadequate but in certain respects perverse. Ultimately, a robust regime of ex ante frontier-AI regulation—under which government institutions or credibly neutral third-party experts are empowered to ensure that AI developers are properly investigating and mitigating the risks of foundation-model development—is urgently required.

Introduction

Many of the machine learning methods that lie at the heart of modern AI development have existed in some form for several decades, but recent technological breakthroughs have enormously increased the scope of what these methods can achieve.[1] Over the last several years, AI developers have discovered novel algorithmic architectures and powerful new training techniques,[2] enabling them to leverage vast amounts of data and computational power. As a result, today's most powerful AI systems are far more capable than the most advanced systems from a couple of years ago—and their capabilities are rapidly advancing beyond human-level in a number of fields.[3] The frontier of AI development is dominated by a handful of AI companies (such as OpenAI, Anthropic, Google DeepMind, Meta, and xAI), several of which work in close partnership with a small group of cloud-compute providers (such as Amazon and Microsoft).[4] Together, these companies are spending hundreds of billions of dollars to develop and deploy the next generation of frontier AI systems at breakneck speed.[5]

Frontier AI systems derive their abilities from the *foundation models* that lie at their core.[6] A "model" is a complex mathematical function which maps inputs to outputs. Until quite recently, the most commercially significant AI systems were built on *narrow models,* mathematical

1. *See generally* Wayne Xin Zhao et al., *A Survey of Large Language Models,* ARXIV (Mar. 18, 2026), https://arxiv.org/pdf/2303.18223 [https://perma.cc/4R7R-9KAE] (discussing the development of these machine learning methods); Shervin Minaee et al., *Large Language Models: A Survey,* ARXIV (Mar. 23, 2025), https://arxiv.org/pdf/2402.06196 [https://perma.cc/RN74-ZWWY] (same).

2. *See* FT Visual Storytelling Team & Madhumita Murgia, *Generative AI Exists Because of the Transformer*, FIN. TIMES (Sep. 12, 2023), https://ig.ft.com/generative-ai [https://perma.cc/HRG2-X46D].

3. Sha Sajadieh et al., *The AI Index 2026 Annual Report*, STAN. INST. FOR HUMAN-CENTERED A.I. 9 (Apr. 2026), https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf [https://perma.cc/5B5X-YQKV] ("AI capability is not plateauing. It is accelerating and reaching more people than ever. Industry produced over 90% of notable frontier models in 2025, and several of those models now meet or exceed human baselines on PhD-level science questions, multimodal reasoning, and competition mathematics. On a key coding benchmark—SWE-bench Verified—performance rose from 60% to near 100% of meeting the human baseline in a single year."); *id.* at 71 ("AI capability is outpacing the benchmarks designed to measure it, and surpassing human-level performance.").

4. *See Data on Notable AI Models*, EPOCH AI, https://epoch.ai/data/notable-ai-models [https://perma.cc/A7QA-DSFS] (presenting a detailed database tracking factors that drive machine learning); Off. of Tech. Staff, *Partnerships Between Cloud Service Providers and AI Developers: FTC Staff Report on AI Partnerships & Investments 6(b) Study*, FED. TRADE COMM'N (Jan. 2025), https://www.ftc.gov/system/files/ftc_gov/pdf/p246201_aipartnerships6breport_redacted_0.pdf [https://perma.cc/L6V4-NH8P].

5. *See, e.g.*, Vlad Savov, *US Big Tech Ratchets Up AI Spending Past $700 Billion This Year*, BLOOMBERG (Apr. 30, 2026), https://www.bloomberg.com/news/articles/2026-04-30/us-big-tech-ratchets-up-ai-spending-past-700-billion-this-year [https://perma.cc/J56A-RJR7].

6. *See generally* Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, ARXIV (July 12, 2022), https://arxiv.org/pdf/2108.07258 [https://perma.cc/N7MM-LG7W] (discussing foundation models' capabilities, design, and significance).

functions derived from (or "trained on") domain-specific data sets.[7] Just like the data from which these models are derived, the capabilities of narrow models are domain-specific. A narrow model might operate a self-driving car or diagnose diseases, but the same narrow model will not do both.

By contrast, foundation models are initially trained on extremely large data sets, containing much of the text on the internet.[8] After that, their capabilities are enhanced through various other techniques, such as powerful forms of reinforcement learning and additional training on domain-specific data ("fine-tuning").[9] As a consequence of this process, foundation models are among the most general-purpose technologies ever devised. Put another way, they are extremely *polymathic*: they provide services across a wide range of commercial, social, military, and political domains.[10] Foundation models are also *protean.* Not only can they function as chatbots and simple tools, but they are increasingly capable of functioning as *agents* that form and execute complex plans of action in the digital world. Foundation models have started to serve as software programmers, science researchers, generalist research assistants, and customer-service representatives.[11]

Today's foundation models remain unreliable in many respects, but their reliability and power are increasing at an extremely rapid rate.[12] At some tasks, such as finding and exploiting software vulnerabilities, foundation models have arguably surpassed the capabilities of "all but the

7. *See* Elliot Jones, *What Is a Foundation Model?,* ADA LOVELACE INST. (July 17, 2023), https://www.adalovelaceinstitute.org/resource/foundation-models-explainer [https://perma.cc/4QHL-ARKC] (describing the unique features and purpose of foundation models).

8. *See* Section I.A, *infra.*

9. *See infra* Sections I.B, I.C.

10. *See* Lewis Hammond et al., *Multi-Agent Risks from Advanced AI*, ARXIV 4 (Feb. 19, 2025), https://arxiv.org/pdf/2502.14143 [https://perma.cc/5Z63-5LJS] (noting that AI agents are already responsible for managing large financial portfolios and recommending military actions, among other tasks, and that they will soon be utilized in energy management, transport networks, and other critical infrastructure).

11. *See, e.g.*, *Introducing Codex*, OPENAI (May 16, 2025), https://openai.com/index/introducing-codex [https://perma.cc/ZMJ2-Q7R7]; *Introducing Claude 4*, ANTHROPIC (May 22, 2025), https://www.anthropic.com/news/claude-4 [https://perma.cc/D3ZX-92FU] (describing Claude 4 as "the world's best coding model"); Juraj Gottweis & Vivek Natarajan, *Accelerating Scientific Breakthroughs with an AI Co-Scientist*, GOOGLE RSCH. (Feb. 19, 2025), https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist [https://perma.cc/EN8T-CMZA]; *Meet Your AI Assistant*, LINDY, https://www.lindy.ai [https://perma.cc/RQ25-XNTY]; *The AI Code Editor*, CURSOR, https://www.cursor.com/en [https://perma.cc/ZV8V-6ENM]; *infra* Section I.C.

12. *See Frontier AI Trends Report*, UK AI SEC. INST. 12-22 (Dec. 2025), https://aisi.s3.eu-west-2.amazonaws.com/Frontier+AI+Trends+Report+-+AI+Security+Institute.pdf [https://perma.cc/S4NZ-3YX8] (documenting rapid capability gains in safety-critical domains*); see also Machine Learning Trends*, EPOCH AI (Dec. 9, 2025), https://epoch.ai/trends [https://perma.cc/S6S9-HVP6] (discussing how the underlying technical determinants of frontier-model capability, such as training compute, have changed).

most skilled humans."[13] For example, Anthropic claims that one of its newest models "has already found thousands of high-severity vulnerabilities in every major operating system and web browser."[14] Some of the field's leading researchers and executives believe that artificial agents more capable than any human in substantially *all* fields will arrive within the decade.[15]

It is entirely possible that these forecasts are overly optimistic and that the rapid pace of frontier AI development will falter. Even if it does, however, the history of AI development suggests that progress will probably resume before too long.[16] At some point in the future, perhaps very soon, we will probably inhabit a digital world filled with foundation models functioning as sophisticated artificial agents and speakers, matching or outstripping human capabilities in many domains. Determining how our legal institutions ought to handle this seismic technological development is a pressing task. Perhaps the oldest of the relevant legal regimes—one which automatically governs new technological developments even before statutes and regulations are passed to address them—is the common law of torts.

This Article provides an in-depth normative, conceptual, and doctrinal examination of tort liability for foundation-model development and release. To date, legal scholarship on AI liability has focused mostly on narrow models, such as those that power self-driving cars and medical devices.[17] In the last few years, however, state and federal courts have

13. *Project Glasswing: Securing Critical Software for the AI Era*, ANTHROPIC (Apr. 7, 2026), https://www.anthropic.com/glasswing [https://perma.cc/8TKB-NN33].

14. *Id.* (emphasis omitted).

15. *See* Benj Edwards, *Anthropic Chief Says AI Could Surpass "Almost All Humans at Almost Everything" Shortly After 2027*, ARS TECHNICA (Jan. 22, 2025), https://arstechnica.com/ai/2025/01/anthropic-chief-says-ai-could-surpass-almost-all-humans-at-almost-everything-shortly-after-2027 [https://perma.cc/FVT3-D6QB] (interviewing Anthropic CEO Dario Amodei on the rapidity of AI model advancement).

16. *See generally* Amirhosein Toosi et al., *A Brief History of AI: How to Prevent Another Winter (a Critical Review)*, ARXIV 4-11 (Sep. 3, 2021), https://arxiv.org/pdf/2109.01517 [https://perma.cc/LN9Y-Y8ZT] (discussing previous starts and stops in AI development).

17. Examples abound. *See generally* Mark A. Geistfeld, *A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation,* 105 CALIF. L. REV. 1611, 1615 (2017) (discussing self-driving cars); Kenneth S. Abraham & Robert L. Rabin, *Automated Vehicles and Manufacturer Responsibility for Accidents: A New Legal Regime for a New Era*, 105 VA. L. REV. 127 (2019) (same); Catherine M. Sharkey, *A Products Liability Framework for AI*, 25 COLUM. SCI. & TECH. L. REV. 240 (2024) (discussing AI-powered medical devices); Charlotte A. Tschider, *Medical Device Artificial Intelligence: The New Tort Frontier*, 46 BYU L. REV. 1551 (2021) (same); Jane R. Bambauer, *Dr. Robot*, 51 U.C. DAVIS L. REV. 383 (2017) (discussing autonomous medical diagnosticians). When tort scholars have engaged with foundation-model development, they have tended to focus on specific doctrinal issues involving foundation-model speech. *See, e.g.*, Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. FREE SPEECH L. 489, 491-94 (2023); Nina Brown, *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*, 3 J. FREE SPEECH L. 389, 389-92 (2023); Jane Bambauer, *Negligent AI Speech: Some Thoughts About Duty*, 3 J. FREE SPEECH L. 343, 344 (2023) [hereinafter Bambauer, *Negligent AI Speech*]. An older literature explores liability for the acts of artificial agents, but this literature largely predates the advent of foundation models and does not

started to adjudicate the first major batch of tort suits about foundation models. [18] Meanwhile, state and federal legislatures have started to entertain and enact statutes that modify, codify, or augment the tort-liability rules that apply to AI developers, mostly with an eye toward the sorts of tortious harms (e.g., chatbot-induced suicides) that foundation models have allegedly caused.[19]

(therefore) discuss these models' central technical characteristics. *See, e.g.*, Curtis E.A. Karnow, *Liability for Distributed Artificial Intelligences,* 11 BERKELEY TECH. L.J. 147 *passim* (1996). *But see* 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 *passim* (2024) (defending strict liability for catastrophic harms caused by AI models, including foundation models); Gabriel Weil, *Closing the AI Accountability Gap: Strict Liability and Punitive Damages for Advanced Artificial Intelligence*, 106 OR. L. REV. (forthcoming 2027) (manuscript at 47), https://ssrn.com/abstract=4694006 [https://perma.cc/E429-TCEF] (noting a stricter standard might be applied to foundation models than to narrow models).

18. *See, e.g.*, Garcia v. Character Techs., Inc., 785 F. Supp. 3d 1157, 1166-69 (M.D. Fla. 2025) (order on motions to dismiss); Complaint at 1-5, Montoya v. Character Techs., Inc., No. 25-cv-02907 (D. Colo. filed Sep. 15, 2025), https://www.bloomberglaw.com/product/blaw/document/X369E2K11SS93JRCM8M9C8JAFSP [https://perma.cc/ZAF6-38JV]; Complaint at 1-4, Raine v. OpenAI, Inc., No. CGC-25-628528 (Cal. Super. Ct. filed Aug. 26, 2025), https://www.law.berkeley.edu/wp-content/uploads/2025/09/Raine-v-OpenAI.pdf [https://perma.cc/ZN46-MSZJ]; Complaint at 1-2, Shamblin v. OpenAI, Inc., No. 25STCV32382 (Cal. Super. Ct. filed Nov. 6, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/11/SHAMBLIN-v.-OPENAI-Complaint.pdf [https://perma.cc/4TQJ-AKLS]; Complaint at 1-2, Lacey v. OpenAI, Inc., No. CGC-25-630808 (Cal. Super. Ct. filed Nov. 6, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/11/CEDRIC-LACEY-vs.-OPENAI-INC.-COMPLAINT.pdf [https://perma.cc/M73V-2LFE]; Complaint at 1-2, Enneking v. OpenAI, Inc., No. CGC-25-630809 (Cal. Super. Ct. filed Nov. 6, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/11/ENNEKING-vs.-OPENAI-INC-COMPLAINT.pdf [https://perma.cc/738V-CJHF]; Complaint at 1-2, Fox v. OpenAI, Inc., No. 25STCV32379 (Cal. Super. Ct. filed Nov. 6, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/11/FOX-v.-OPENAI-Complaint.pdf [https://perma.cc/LLJ5-4Q7D]; Complaint at 1-2, Irwin v. OpenAI, Inc., No. CGC-25-630811 (Cal. Super. Ct. filed Nov. 6, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/11/Irwin-v-OpenAI-COMPLAINT-Nov-6-2025.pdf [https://perma.cc/A5WG-JZAH]; Complaint at 1-2, Madden v. OpenAI, Inc., No. 25STCV32383 (Cal. Super. Ct. filed Nov. 6, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/11/Madden-v.-OpenAI-Complaint.pdf [https://perma.cc/GAG2-Q8FJ]; Complaint at 1-28, Brooks v. OpenAI, Inc., No. 25STCV32386 (Cal. Super. Ct. filed Nov. 6, 2025), https://www.bloomberglaw.com/product/blaw/document/X6Q5Q65LI3580APMRQJ8OTUMGUU/download [https://perma.cc/GVB9-FLDF]; Complaint at 1-5, First Cnty. Bank v. OpenAI Found., No. CGC-25-631477 (Cal. Super. Ct. filed Dec. 11, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/12/FIRST_COUNTY_BANK_vs_OPENAI_FO.pdf [https://perma.cc/8MB9-UJYN]; First Amended Complaint at 1-2, St. Clair v. X.AI Holdings Corp., No. 26-cv-00386 (S.D.N.Y. filed Jan. 15, 2026), https://www.bloomberglaw.com/product/blaw/document/X4VQ7KCBJO89DCPA5GCQQIQFJ57/download [https://perma.cc/YW5M-Z6EB]; *see also* Shweta Watwe, *Character.AI and Google Agree to Settle Teen Chatbot Harm Lawsuits*, BLOOMBERG LAW (Jan. 8, 2026), https://news.bloomberglaw.com/litigation/character-ai-google-agree-to-settle-teen-chatbot-harm-lawsuits [https://perma.cc/UL4X-9YS6] (reporting settlement of Garcia and four related Character.AI cases filed in Florida, Texas, Colorado, and New York).

19. *See, e.g.*, AI-LEAD Act, S. 2937, 119th Cong. (2025) (proposing a federal tort-like private right of action, grounded in products liability principles, against AI developers); Artificial Intelligence: Defenses, Assemb. B. 316, 2025-2026 Leg., Reg. Sess. (Cal. 2025) (codified at CAL. CIV. CODE § 1714.46) (prohibiting defendants from asserting AI autonomy as a defense to civil liability); Companion Chatbots, S.B. 243, 2025-2026 Leg., Reg. Sess. (Cal. 2025) (codified at CAL.

The fundamental question that both judges and legislators must answer, in one posture or another, is what standards of tort liability ought to govern this novel domain. Under existing law, the only sort of tort liability that clearly applies to AI developers is ordinary negligence liability, which governs all activities that pose foreseeable risks of injury to others.[20] Legal scholars, however, have for the most part assumed that negligence liability is normatively ill-suited to AI.[21] So, many scholars have argued that negligence ought to be substantially or entirely displaced by alternative doctrinal frameworks, such as strict liability for abnormally dangerous activities, vicarious liability, or products liability.[22] As a result, the law of negligence's application to AI has been underexplored.[23]

BUS. & PROF. CODE §§ 22601-22606) (creating private right of action for certain injuries from companion-chatbot platforms); S.B. 760, 103rd Leg., Reg. Sess. (Mich. 2025) (proposing a private right of action, including punitive damages, for minors harmed by AI-companion chatbots); Artificial Intelligence Bill of Rights, S.B. 482, 2026 Leg., Reg. Sess. (Fla. 2026) (proposing allowing suits by minors and guardians for up to $10,000 per violation by companion-chatbot operators); *see also* Transparency in Frontier Artificial Intelligence Act, ch. 138, § 1(p), 2025 Cal. Legis. Serv. Ch. 138 (West) (mandating safety frameworks and incident reporting for frontier-model developers, while recognizing in findings that "collective safety will depend in part on frontier developers taking due care in their development and deployment of frontier models proportional to the scale of the foreseeable risks"); Responsible AI Safety and Education Act, 2025 N.Y. Laws 699 (codified at N.Y. GEN. BUS. LAW §§ 1420-1425 (McKinney 2026)) (same).

20. Certain jurisdictions may determine that (certain) AI models are products, subjecting their developers (and others in the AI supply chain) to products liability. But the question is open; to date, no state court has clearly resolved whether software is a sort of "service," subject to ordinary negligence liability, or a sort of "product" subject to specialized product liability doctrine (some of which is strict, rather than fault-oriented, in character). *See generally* Asaf Lubin, *On Software Bugs and Legal Bugs: Product Liability in the Age of Code*, 100 IND. L.J. 1891 (2025) (surveying the approaches of different states to this question).

21. *See* Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. 1315, 1318 (2020) ("The large and growing body of scholarship on AI and tort law has mostly set aside discussions of negligence . . . .").

22. *See, e.g.*, Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 395 (2016) (arguing that "strict liability [should] ordinarily attach" to AI that has not been certified as safe by a dedicated regulatory agency); Henson, *supra* note 17, at 362-72; Weil, *supra* note 17 (manuscript at 45-50, 54-57); Abraham & Rabin, *supra* note 17, at 145-49, 156-64; Anat Lior, *AI Strict Liability Vis-À-Vis AI Monopolization*, 22 COLUM. SCI. & TECH. L. REV. 90, 94-97, 125-26 (2020); Christiane Wendehorst, *Strict Liability for AI and Other Emerging Technologies*, 11 J. EUR. TORT L. 150, 165-70, 177-80 (2020); *see also* Jin Yoshikawa, *Sharing the Costs of Artificial Intelligence: Universal No-Fault Social Insurance for Personal Injuries*, 21 VAND. J. ENT. & TECH. L. 1155, 1178-80 (2019) (proposing universal no-fault social insurance in place of most tort litigation for covered AI injuries).

23. *But see* Ian Ayres & Jack Balkin, *The Law of AI Is the Law of Risky Agents Without Intentions*, U. CHI. L. REV. ONLINE *8 (2024), https://lawreview.uchicago.edu/sites/default/files/2024-11/Ayres_Balkin_Law%20of%20Risky%20Agents.pdf [https://perma.cc/V3LY-TW8E] (arguing that "a negligence-based approach offers the flexibility needed to respond to emerging challenges while holding designers and users liable for failure to exercise reasonable care"); Chinmayi Sharma, *AI's Hippocratic Oath*, 102 WASH. U. L. REV. 1101, 1142-59 (2025) (suggesting a professional standard of care for AI developers, on the model of negligence liability for medical professionals); Bryan H. Choi, *AI Malpractice*, 73 DEPAUL L. REV. 301, 306 (2024) (same); Kenneth S. Abraham & Catherine M. Sharkey, *Untangling AI Liability*, 115 CALIF. L. REV. (forthcoming 2027) (manuscript at 7-8, 14-15) (arguing that some harms should be subject to products liability, strict liability, or no liability while others should be subject to negligence).

This Article provides a qualified defense of the tort of negligence as the doctrinal foundation for the tort system's governance of the development and release of foundation models. As applied to foundation-model development, the aforementioned doctrinal alternatives (strict liability for abnormally dangerous activities, vicarious liability, and products liability) are subject to a range of underappreciated normative, conceptual, and institutional problems and limitations.[24] These are diverse in character, but many of them have a common core: the particular assumptions that underlie these specialized doctrinal regimes founder on the polymathic and protean character of foundation models. By contrast, generality and flexibility are the defining characteristics of the law of negligence and the broad duty to take reasonable care that lies at its core.

The breadth and flexibility of the negligence tort will allow it to provide redress for the range of important harms that foundation-model developers may cause—including harms caused via routes that specialized forms of liability will struggle to cover, such as risky internal deployments, failures to adequately secure models against theft or misappropriation, open-source releases, customized releases, and careless entrustments of highly dangerous (but nondefective) models.[25] Moreover, the tort of negligence can register the normative significance of the externalized benefits that foundation models will foreseeably produce. Stricter forms of liability, such as strict liability for abnormally dangerous activities, may unduly suppress these benefits by distorting developers' choices about whether and how to make models' powerful functionalities available.[26]

That said, exclusive reliance on negligence principles to determine the liability of foundation-model developers does face a few significant

24. For the most part, this Article focuses on normative considerations involving the deterrence and mitigation of harm and the provision of social benefits, rather than considerations of interpersonal justice between plaintiff and defendant. That is not because the latter considerations are normatively or interpretively insignificant. To the contrary, I believe that such ideas lie at the core of tort law's doctrinal architecture (even if "tort law also leverages this moralized doctrinal apparatus to serve larger societal ends"). *See* Ketan Ramakrishnan, *What Is a Tort?,* 139 HARV. L. REV. 1011, 1015 n.26 (2026). Rather, this Article focuses on societal normative considerations for three reasons. First, when the tort system is governing the imposition of truly gigantic risks of harm, such as risks of biological catastrophe, it is plausible that aggregate harm prevention and benefit provision swamp the normative significance of interpersonal justice between litigants. Enormous risks of this kind are among the most salient risks that frontier-AI development may pose. Second, both tort doctrine and ordinary moral thought are deeply torn (and perhaps internally inconsistent) as to when interpersonal justice favors fault-based liability over strict liability (and vice versa). *Id.* at 1067-76. And it is the choice between (various forms of) fault-based and strict liability that is (largely) this Article's normative concern. Finally, even if one assumes that the sole affirmative *goal* of tort law should be to effectuate interpersonal justice between defendant and plaintiff, it is highly plausible—even from within the auspices of a purely rights-based moral theory—that a tort-liability regime may be normatively objectionable because it imposes large harmful *externalities* on third parties, such as by suppressing access to widely beneficial defensive tools. *Cf.* Sandy Steel, *Deterrence in Private Law*, *in* PRIVATE LAW AND PRACTICAL REASON 123, 134-35 (Haris Psarras & Sandy Steel eds., 2023) (discussing overdeterrence). Much of the argument in Part II can be understood along these lines.

25. *See infra* Sections II.A.1, II.B.1, and II.C.1.

26. *See infra* Section II.A.1.

objections. The first and most novel is the *remedial black-hole* problem.[27] When applied to AI models, existing negligence doctrine provides no basis for relief in certain situations that seem to clearly call for it. Suppose, for example, that as the result of an AI developer's negligence, its foundation-model sales agent malfunctions and defrauds a consumer. Or suppose that a negligent AI developer releases a model that foreseeably interacts in a highly sexualized manner with an underage user, subjecting them to severe emotional distress.[28] Without further judicial development, black-letter negligence doctrine will not hold the developer (or any other party) liable in either case, because the developer has no duty of care: negligence law imposes a duty to take care against inflicting economic loss or emotional injury only in narrow circumstances, and neither of the preceding cases qualifies.[29]

The remedial black-hole problem might be addressed by subjecting AI developers to vicarious liability for the "torts" of their AI models, as some scholars have proposed. But such a remodeling of tort doctrine faces serious conceptual and normative difficulties when applied to foundation models.[30] Classifying AI models as products subject to products liability may be a more promising solution. In both of the preceding hypothetical cases, after all, the model's performance would seem to contravene legitimate safety expectations of consumers, and at least one important form of products liability looks to such expectations to determine when a product is defective. But the limited scope of products liability means that it will not provide a general solution to the issue: even assuming that AI models are properly categorized as products, products liability doctrine will not cover the many cases where a model may cause harm without commercial sale or distribution (e.g., internal deployments, weight theft or misappropriation, and many open-source releases).

Thus, this Article suggests a different solution, one that operates *within* the law of negligence: in appropriate cases, courts should enrich the content of the duty of care in negligence by drawing from intentional-tort doctrine, criminal-law doctrine, and other bodies of law (such as contract, estoppel, and fiduciary law).[31] That is, courts should recognize that AI developers have a duty to take care against releasing artificial agents that perform actions that would, if these agents were legal persons, violate legal prohibitions drawn from the law of intentional torts and the criminal law. An alternative (and largely complementary) strategy would be to enrich

27. See *infra* Sections II.C.3, III.A.

28. The latter sort of scenario may already have occurred, possibly many times. *See* Caitlin Gibson, *Her Daughter Was Unraveling, and She Didn't Know Why. Then She Found the AI Chat Logs.*, WASH. POST. (Dec. 23, 2025), https://www.washingtonpost.com/lifestyle/2025/12/23/children-teens-ai-chatbot-companion [https://perma.cc/L937-Z8VU].

29. *See infra* Sections II.C.3, III.A.

30. *See infra* Section II.C.

31. *Infra* Section III.A.

duty analysis in negligence by drawing upon the idea of safety expectations, which currently plays a greater role in products liability than in negligence: courts could hold that AI developers have a duty to take care against causing economic or emotional injury by releasing a model that behaves in a manner that contravenes certain clear safety expectations of consumers or third parties.

The other issue, which we might call the *proof-of-breach* problem, is not particular to the context of foundation models or AI—but the advent of foundation models may require this problem to be solved in novel ways. Suppose that a chatbot actively encourages an underage user to commit suicide,[32] or validates a clearly psychotic user's evidently baseless and delusional belief that his mother is conspiring to kill him, thus leading him to kill his mother.[33] It seems plausible, in such cases, that the injured plaintiff ought to be legally entitled to obtain relief even without providing any particular evidence (beyond the bare fact of the model's behavior) that the developer has failed to take reasonable care.

To be clear, the issue here is not that the proof-of-breach problem is likely to leave such a plaintiff without a viable negligence claim. The very egregiousness of the model's behavior, in such cases, suggests that the injured plaintiff will likely be able to discover sufficient evidence of the developer's negligence to satisfy her burden of production on breach and get to the jury (supposing that she can meet her burden of production on injury and causation).[34] Rather, the issue is that imposing this evidentiary burden may subject the plaintiff to additional litigation costs and legal uncertainty—decreasing the expected value of the settlements that she can obtain and possibly deterring her from bringing suit at all.

That injured tort plaintiffs might sometimes face unwarranted evidentiary burdens is, of course, a general issue; a concern of this kind partly motivated the creation and development of strict products liability.[35] Thus, one obvious solution to the proof-of-breach problem in this context would be to allow certain plaintiffs injured by AI models to proceed against developers or deployers not only on a theory of negligence but also on a

32. *Cf.* Complaint at 1-5, Raine v. OpenAI, Inc., No. CGC-25-628528 (Cal. Super. Ct. filed Aug. 26, 2025), https://www.law.berkeley.edu/wp-content/uploads/2025/09/Raine-v-OpenAI.pdf [https://perma.cc/ZN46-MSZJ] (presenting similar facts).

33. *Cf.* Complaint at 1-5, First Cnty. Bank v. OpenAI Found., No. CGC-25-631477 (Cal. Super. Ct. filed Dec. 11, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/12/FIRST_COUNTY_BANK_vs_OPENAI_FO.pdf [https://perma.cc/8MB9-UJYN] (making similar allegations).

34. Some initial evidence for this prediction is provided by the trial court's order refusing to grant the defense's motion to dismiss in *Garcia v. Character Technologies, Inc. See* 785 F. Supp. 3d 1157, 1181 (M.D. Fla. 2025).

35. *See, e.g.*, Escola v. Coca Cola Bottling Co., 150 P.2d 436, 441 (Cal. 1944) (Traynor, J., concurring) ("An injured person, however, is not ordinarily in a position to . . . identify the cause of the defect, for he can hardly be familiar with the manufacturing process as the manufacturer himself is."); MARK A. GEISTFELD, PRINCIPLES OF PRODUCTS LIABILITY 27-29 (3d ed. 2020) (developing such an evidentiary rationale for strict products liability at greater length).

theory of strict products liability. But even if state courts decide to classify AI models as products—and some states may not—this solution has significant limitations. For one thing, as already noted, products liability has a limited scope; it will not cover many sorts of tortious injuries that foundation models might inflict, through routes such as internal deployments.[36] Moreover, products liability's focus on completed products over procedures, and consumer interests over the interests of all affected parties, may lead the products liability framework to distort the structure of juror deliberation and decision making in certain cases.[37] Reflecting on these potential doctrinal distortions suggests that—while extending products liability to AI models *may* be warranted, on balance—the case for doing so is more uneasy than many torts scholars (and the first court to rule on the matter) have supposed. In any event, the law of AI liability cannot adequately address the proof-of-breach problem simply by extending products liability to AI models, given products liability's limited scope.

This Article thus suggests a solution to the proof-of-breach problem that—like the Article's solution to the remedial black-hole problem—operates within the boundaries of negligence doctrine itself. A plaintiff who establishes that an AI model has behaved in a clearly flawed way—that the model has, in an intuitive sense, *malfunctioned*—should be allowed to thereby satisfy her burden of production on breach (along the lines of the permissive inference allowed by doctrines such as res ipsa *loquitur* in negligence and the malfunction doctrine in products liability). Put another way, the court should allow the jury to find for the plaintiff even if she does not provide expert testimony or any other evidence of breach beyond the bare fact of the malfunction. Under a more aggressive version of this solution, a plaintiff who makes such a showing would shift the burden of persuasion on breach to the defendant.[38] But even the less aggressive version ought substantially to alleviate the proof-of-breach problem.

---

36. See *infra* Section II.B.1.

37. See *infra* Section II.B.2.

38. Tort scholars who believe that the development of strict products liability was a mistake and ought to be abandoned but who nevertheless wish to alleviate the burden of proof for a plaintiff injured by a manufacturing defect have made similar doctrinal proposals with respect to injuries occasioned by manufacturing defects. *See, e.g.*, William Powers, Jr., *A Modest Proposal to Abandon Strict Products Liability*, 1991 U. ILL. L. REV. 639, 641 ("[M]y proposal is (1) to abandon strict products liability as a separate body of law, (2) to rely solely on negligence in product cases, but (3) to have a special rule within negligence to govern manufacturing defects. This special rule for manufacturing defects might provide that a manufacturer's violation of its own specifications raises a permissive inference of negligence, meaning that a court could not direct a verdict against a plaintiff for failing to prove what caused the flaw. It might provide that a manufacturer's violation of its own specifications raises a presumption of negligence, rebuttable only by a defendant's own evidence of reasonable care." (internal footnotes omitted)). The proposal is elegant and plausible, but it cannot simply be ported over to deal with the proof-of-breach problem in the case of foundation models with egregious flaws, because it is probably not appropriate to characterize such flaws as manufacturing defects. *See infra* note 190 and accompanying text.

Call this the *behavioral malfunction* doctrine. There are various ways in which the operative idea of a behavioral malfunction might be given greater doctrinal specificity: by reference to the fact that the model's behavior would, if committed by a human, constitute an intentional tort or crime; by reference to the fact that the model's behavior violates clear and specific societal safety expectations;[39] or by reference to the fact that the model's behavior contravenes the developer's own specific intentions regarding the model's desired behavior (or specific intentions that it would be negligent for the developer *not* to have or act on). Alternatively, the operative idea of a malfunction might be left doctrinally unspecified, at least to begin—much as the central notion of a defect was largely left unspecified in the early stages of American products liability law.[40] In any of these forms, the behavioral malfunction doctrine would represent an incremental doctrinal development—one that extends well-established principles of existing negligence law in an only moderately creative manner to reflect novel features of foundation models (such as their ability to act in quasi-unlawful ways).

Before developing these doctrinal and normative arguments, this Article begins by providing a comprehensive but accessible overview of how foundation models are developed and released. Outlining this process illuminates some of the central risks that foundation models pose, and the strategies that developers use to mitigate them. Some of these risks—such as causing certain users to commit suicide and homicide, enabling bad actors to create nonconsensual deepfake pornography, and facilitating cyberattacks on important public infrastructure—may already be materializing.[41] Soon, they may well materialize in even graver form. Much of the AI industry itself, perhaps partly because of its exposure to tort liability, is taking some of these risks seriously. Anthropic, for example, delayed the public release of a powerful recent model because of the

39. On why the relevant expectations should be those of society, rather than consumers, see *infra* Section II.B.2.

40. *Cf.* Michael D. Green, *The Unappreciated Congruity of the Second and Third Torts Restatements on Design Defects*, 74 BROOK. L. REV. 807, 807 (2009) (noting this early ambiguity).

41. *See, e.g.*, Complaint, ¶¶ 134-94, Garcia v. Character Techs., Inc., 785 F. Supp. 3d 1157 (M.D. Fla. 2025); Complaint at 1-5, Raine v. OpenAI, Inc., No. CGC-25-628528 (Cal. Super. Ct. filed Aug. 26, 2025), https://www.law.berkeley.edu/wp-content/uploads/2025/09/Raine-v-OpenAI.pdf [https://perma.cc/ZN46-MSZJ]; Complaint at 1-5, 10-32, A.F. *ex rel.* J.F. v. Character Techs., Inc., No. 24-cv-01014 (E.D. Tex. filed Dec. 9, 2024), https://storage.courtlistener.com/recap/gov.uscourts.txed.234704/gov.uscourts.txed.234704.1.0_2.pdf [https://perma.cc/ZP6Q-73CN]; Complaint at 1-5, First Cnty. Bank v. OpenAI Found., No. CGC-25-631477 (Cal. Super. Ct. filed Dec. 11, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/12/FIRST_COUNTY_BANK_vs_OPENAI_FO.pdf [https://perma.cc/8MB9-UJYN]; Nick Robins-Early, *US Man Used AI to Generate 13,000 Child Sexual Abuse Pictures, FBI Alleges*, GUARDIAN (May 21, 2024), https://www.theguardian.com/technology/article/2024/may/21/child-sexual-abuse-material-artificial-intelligence-arrest [https://perma.cc/K26K-MTVE]; Jordan Robertson & Katrina Manson, *Hacker Used Anthropic's Claude to Steal Sensitive Mexican Data*, BLOOMBERG (Feb. 25, 2026), https://www.bloomberg.com/news/articles/2026-02-25/hacker-used-anthropic-s-claude-to-steal-sensitive-mexican-data [https://perma.cc/SQR9-NCVA].

model's extraordinary cyberoffensive capabilities.[42] And recently, a rogue OpenAI model escaped from a contained testing environment and conducted a sophisticated autonomous hacking operation against another company's online database.[43]

Other sorts of risks—such as the risk that foundation models will enable malicious actors to successfully engage in biological terrorism[44]—have not yet materialized (for all the public record discloses). But these risks are quickly growing in magnitude and severity. For example, in February 2025, OpenAI reported that its recent biological-risk evaluations suggest that its "models are on the cusp of being able to meaningfully help novices create known biological threats."[45] In June 2026, OpenAI reported that one of its models has met that threshold.[46] Other recent work suggests that frontier models are making substantial gains in providing virology assistance to experts.[47]

Foundation-model development is currently subject to anemic and haphazard regulatory oversight and restriction.[48] At least until more comprehensive and stable mechanisms of *ex ante* governance are implemented, it is largely the law of torts—which governs a wide swath of risky behavior by default—that will govern many of the risks and harms of developing and releasing frontier AI. Even once substantial *ex ante*

42. *See* Sam Sabin, *Anthropic Holds Mythos Model Due to Hacking Risks,* AXIOS (Apr. 7, 2026), https://www.axios.com/2026/04/07/anthropic-mythos-preview-cybersecurity-risks [https://perma.cc/7U7A-PDZ5] ("Anthropic is so worried about the damage . . . [its own model] could cause that it's refusing to release it publicly until there are safeguards to control its most dangerous capabilities.").

43. *See* Hugo Larcher, Adrien Carreira, Raphael G & Christophe Rannou, *Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident*, HUGGING FACE (July 27, 2026), https://huggingface.co/blog/agent-intrusion-technical-timeline [https://perma.cc/A9B6-D6DF].

44. *See* Christopher A. Mouton, Caleb Lucas & Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks: A Red-Team Approach*, RAND (2023), https://www.rand.org/pubs/research_reports/RRA2977-1.html [https://perma.cc/VY6T-H6RS].

45. *See Deep Research System Card*, OPENAI 17 (Feb. 25, 2025), https://cdn.openai.com/deep-research-system-card.pdf [https://perma.cc/T97A-JA8M].

46 . *See GPT-Rosalind-5.5 System Card,* OPENAI 2 (June 3, 2026), https://deploymentsafety.openai.com/gpt-rosalind-5-5 [https://perma.cc/2Y69-NP8B].

47. *See* Jasper Götting et al., *Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark*, ARXIV 1, 8-10 (Apr. 21, 2025), https://arxiv.org/pdf/2504.16137 [https://perma.cc/K3NV-7GGL].

48. The development of certain narrow models, such as those used in medical devices and self-driving cars, is more heavily regulated. *See, e.g.*, Kavitha Palaniappan, Elaine Yan Ting Lin & Silke Vogel, *Global Regulatory Frameworks for the Use of Artificial Intelligence (AI) in the Healthcare Services Sector,* 12 HEALTHCARE art. no. 562, at 5-9 (2024); *Self-Driving Vehicles Enacted Legislation*, NAT'L CONF. OF STATE LEGISLATURES (Feb. 18, 2020), https://www.ncsl.org/transportation/autonomous-vehicles [https://perma.cc/92X8-G6PL] (discussing states enacting legislation related to autonomous vehicles). Foundation-model development is also more regulated in the EU, where it is governed by the EU AI Act. *See* Innocenzo Genna, *The Regulation of Foundation Models in the EU AI Act*, INT'L BAR ASS'N (Apr. 12, 2024), https://www.ibanet.org/the-regulation-of-foundation-models-in-the-eu-ai-act [https://perma.cc/EB9A-ET2M]. The efficacy of this regulatory regime remains unclear. It is widely rumored in the AI industry, for example, that at least one frontier-AI developer has declined to sign on to the draft AI Act Code of Practice.

regulation is enacted, risks such as industry capture and limited regulatory competence—risks quite salient in this domain given the dense concentration of top technical expertise within the frontier-AI companies—argue against dispensing entirely with the *ex post* backstop provided by the tort system. Moreover, there is some reason to believe that substantial administrative regulation, when it arrives, will partly build on common-law concepts and liability rules.[49] It is important to get these liability rules right.

That said, in exploring the normative and doctrinal choices between alternative liability rules, we will encounter important limitations of the tort system as an instrument of AI governance. And we will observe that the tort system itself generates certain perverse incentives for investigating and disclosing information about risks and harms. That is especially disturbing in this context, given how heavily our society is relying on frontier-AI companies themselves in order to understand the nature and significance of the risks that frontier-AI development may pose. A well-designed system of *ex ante* regulation could mitigate and counteract these perverse incentives. Thus, as this Article's Conclusion argues, frontier-AI governance urgently requires new regulatory institutions that can oversee the investigation, disclosure, and mitigation of risks *ex ante*—not just *ex post* liability rules that penalize risk imposition after it materializes in harm.

I. The Foundation-Model Pipeline

This Article seeks to clarify how tort law ought to govern foundation-model development and release, and to address tort's prospects, limitations, and pitfalls as a regulatory instrument. Both tasks require understanding the nature of the conduct at issue. Thus, this Part provides an overview of the process by which foundation-model developers train and test foundation models, reduce or inhibit their potentially dangerous behavior, equip them with safeguards against misuse, deploy them internally, and release them to the public. It also explains how downstream developers or "operators" (i.e., other firms or individuals) can modify and augment foundation models.[50] Along the way, it highlights some of the

49. Some early attempts to regulate foundation-model chatbots have sought to mitigate risk by imposing tort-like duties of care upon AI developers. *See, e.g.*, Jason M. Loring & Graham H. Ryan, *The TRUMP AMERICA AI Act: Federal Preemption Meets Comprehensive Regulation*, JONES WALKER: AI L. & POL'Y NAVIGATOR (Jan. 23, 2026), https://www.joneswalker.com/en/insights/blogs/ai-law-blog/the-trump-america-ai-act-federal-preemption-meets-comprehensive-regulation.html [https://perma.cc/9GXH-HBZV] ("The [TRUMP AMERICA AI Act] imposes a duty of care on AI developers to 'prevent and mitigate foreseeable harm to users,' enforceable by the Federal Trade Commission (FTC) through rulemaking authority.").

50. *See Claude's Constitution*, ANTHROPIC, https://www.anthropic.com/constitution [https://perma.cc/3VPU-NG7A] ("We use the term 'principals' to refer to those whose instructions

major forms of potentially tortious malfunction and misuse. By the end of this Part, readers should have a clear picture of what, exactly, the tort system must govern when it addresses frontier-AI development.

A. Training the Baseline Model

Recall that a "model" is just a mathematical function which maps inputs to outputs.[51] The simplest models (e.g., $y = 3x + 1$) are familiar to anyone who has studied elementary algebra. In traditional software programming, human coders manually specify such functions. A traditional software program is thus the product of many conscious human design choices. Similarly, early forms of artificial intelligence were consciously designed by, and intelligible to, human coders.[52]

By contrast, a modern AI model is largely opaque to human understanding. It consists of a vast amount of numbers—often called the model's *weights* or *parameters*—arrayed in the form of many successive matrices.[53] Inputs into the model are converted into mathematical form and multiplied by this array of matrices in order to produce outputs, which are then converted out of mathematical form in order to yield a final output (such as a natural-language sentence). So, for example, a foundation model might take as an input the first few words in a sentence, and yield, as its output, a prediction of the next word which that sentence likely contains. AI models are sometimes called *neural networks* because their high-level architecture and functions are broadly structurally analogous to biological neural networks.[54]

How are such models made? The first major step in the training process is often called *pre-training*,[55] which results in a *pre-trained* or

Claude should give weight to and who it should act on behalf of, such as those developing on Anthropic's platform (operators) and users interacting with those platforms (users).").

51. *See generally* Andreas Stöffelbauer, *How Large Language Models Work*, MEDIUM (2023), https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f [https://perma.cc/KLT2-2U5B] (describing the inputs involved in creating LLMs); DANIEL JURAFSKY & JAMES H. MARTIN, SPEECH AND LANGUAGE PROCESSING (3d ed. draft 2026), https://web.stanford.edu/~jurafsky/slp3/ed3book_jan26.pdf [https://perma.cc/VMG9-KMPS] (providing an overview of modern machine learning); AARON COURVILLE, IAN GOODFELLOW & YOSHUA BENGIO, DEEP LEARNING (2015) (surveying the field of machine learning and deep learning).

52. *See* STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH § 1.3, at 16 (3d ed. 2010) (describing agents as capable of making rational decisions based on what they believe or want).

53. Elizabeth Seger et al., *Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives*, ARXIV 9 (Sep. 29, 2023), https://arxiv.org/pdf/2311.09227 [https://perma.cc/T57J-BJSJ].

54. *See* Fangfang Lee, *What Is a Neural Network?*, IBM, https://www.ibm.com/topics/neural-networks [https://perma.cc/4RCZ-V9F9] (describing how neural networks operate).

55. JURAFSKY & MARTIN, *supra* note 51, at 120-46.

baseline model.[56] Here is how pre-training works, at a relatively high level of abstraction.[57]

Data Curation and Creation: Training any AI model requires a data set that contains some significant pattern or patterns of interest—for example, lots of different natural-language sentences—which the model will be trained to "fit."[58] Foundation models are initially trained on a data set that includes a "vast amount[]" of text, largely extracted from the internet.[59] It is a live possibility that the training data available on the internet (and through other traditional sources) may soon run out.[60] Perhaps it already has.[61] Thus foundation-model developers are increasingly *creating* new training data, including "synthetic data" generated by foundation models themselves.[62] Synthetic data has played a large role in the training of some recent models.[63]

Compute Acquisition: To make use of this data, the training process described in the next few steps requires an enormous amount of computational power, or "compute."[64] Compute is generally provided remotely, through the cloud.[65] A few big companies, such as Amazon,

56. The term "baseline model," which is less common but perhaps more perspicuous, is taken from Peter Henderson, Tatsunori Hashimoto & Mark Lemley, *Where's the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589, 602 (2023).

57. For a deeper dive, see generally ANIL ANANTHASWAMY, WHY MACHINES LEARN: THE ELEGANT MATH BEHIND MODERN AI (2024).

58. JURAFSKY & MARTIN, *supra* note 51, at 44-46.

59. *Id.* at 146-49, 161-62, 169-70.

60. *See generally* Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim & Marius Hobbhahn, *Will We Run Out Of Data? Limits of LLM Scaling Based on Human-Generated Data*, ARXIV (June 4, 2024), https://arxiv.org/pdf/2211.04325 [https://perma.cc/7ZAH-8JPA] (investigating whether the limited availability of public human-generated text will limit LLM scaling).

61. *See* Interview by Mark Penn with Elon Musk, CEO, xAI, at CES 2025 (Jan. 8, 2025) ("We've now exhausted basically the cumulative sum of human knowledge . . . in AI training. That happened basically last year."), *quoted in* Kyle Wiggers, *Elon Musk Agrees That We've Exhausted AI Training Data*, TECHCRUNCH (Jan. 8, 2025), https://techcrunch.com/2025/01/08/elon-musk-agrees-that-weve-exhausted-ai-training-data [https://perma.cc/C6HR-NSXW]; Kylie Robison, *Ilya Sutskever Says the Way AI Is Being Built Is About to Change*, VERGE (Dec. 13, 2024), https://www.theverge.com/2024/12/13/24320811/ilya-sutskever-neurips-2024-talk-pretraining-superintelligence [https://perma.cc/HJ6J-XBJE].

62. *See* Adam Zewe, *In Machine Learning, Synthetic Data Can Offer Real Performance Improvements*, MIT NEWS (Nov. 3, 2022), https://news.mit.edu/2022/synthetic-data-ai-improvements-1103 [https://perma.cc/7H28-CSHH] (describing the role of synthetic data in training baseline models).

63. *See, e.g.*, Aaron Grattafiori et al., *The Llama 3 Herd of Models*, ARXIV, *passim* (July 31, 2024), https://arxiv.org/pdf/2407.21783 [https://perma.cc/ESM7-RRGD].

64. *See* Aadya Gupta & Adarsh Ranjan, *A Primer on Compute*, CARNEGIE ENDOWMENT FOR INT'L PEACE (Apr. 30, 2024), https://carnegieendowment.org/posts/2024/04/a-primer-on-compute [https://perma.cc/JDJ2-ULT2].

65. However, smaller, less powerful AI models can be trained on personal computers or small proprietary computing clusters. *See* Edd Gent, *Powerful AI Can Now Be Trained on a Single Computer*, IEEE SPECTRUM (July 17, 2024), https://spectrum.ieee.org/powerful-ai-can-now-be-trained-on-a-single-computer [https://perma.cc/GP2Y-WDHC]; *Train and Use Your Own Models*, GOOGLE CLOUD, https://cloud.google.com/vertex-ai/docs/training-overview [https://perma.cc/A7SV-PMPT].

Microsoft, Google, and Oracle, dominate this market.[66] In recent years, as the amount of compute required to train frontier foundation models has grown, many AI labs have formed close partnerships with these established cloud providers in order to secure access to sufficiently large quantities of compute.[67] The computational power which these cloud providers offer is ultimately derived from highly specialized computing chips [68] called semiconductors.[69] The manufacturing of these chips is dominated by a handful of companies, such as Nvidia, Intel, and Google.[70]

Architecture Selection: Armed with data and compute, engineers are ready to consider *model architecture*. A model's architecture is, roughly, the high-level structure that organizes its weights, thus dictating how the model turns inputs into outputs. Unlike model weights, model architectures are largely the product of human design choices. The invention of a new model architecture, the transformer, was essential in enabling the advent of modern foundation models, because it let those models process large quantities of data in a more semantically and inferentially sophisticated way than previous architectures allowed.[71] Most frontier foundation models today have architectures that are variations of the transformer; developers continue to vary, adapt, and improve this architecture, resulting in substantial efficiency gains (i.e., reductions in the amount of computational power required to achieve a given level of performance). [72] Meanwhile, new architectures—which may enable

66. *See* Phil Powell & Ian Smalley, *What Is a Hyperscale Data Center?*, IBM (Mar. 21, 2024), https://www.ibm.com/think/topics/hyperscale-data-center [https://perma.cc/X4YH-QHSA].

67. *See, e.g.*, Cecilia Kang & Cade Metz, *Trump Announces $100 Billion A.I. Initiative*, N.Y. TIMES (Jan. 21, 2025), https://www.nytimes.com/2025/01/21/technology/trump-openai-stargate-artificial-intelligence.html [https://perma.cc/MN3G-7EKG] (reporting on President Trump's announcement of a joint venture between OpenAI, SoftBank and Oracle to create $100 billion in computing infrastructure to power artificial intelligence).

68. Chips are the source of computers' computational power (i.e., their ability to perform calculations and process data). *See* Mike Murphy, *What Are Semiconductors?*, IBM (Sep. 14, 2023), https://research.ibm.com/blog/what-are-semiconductors [https://perma.cc/4NHH-VAAS].

69. These specialized chips are informally called semiconductors because their primary material basis is silicon, a sort of semiconductor. *Id.*

70. *See* Kif Leswing, *Nvidia Dominates the AI Chip Market, but There's More Competition than Ever*, CNBC (June 2, 2024), https://www.cnbc.com/2024/06/02/nvidia-dominates-the-ai-chip-market-but-theres-rising-competition-.html [https://perma.cc/72E5-QWY7].

71. For an explanation of how the transformer architecture works and how it enabled modern machine-learning techniques, see JURAFSKY & MARTIN, *supra* note 51, at 173-83.

72. For a discussion of variants of the transformer architecture and alternative architectures, see Minaee et al., *supra* note 1, at 35-36; *see also* Humza Naveed et al., *A Comprehensive Overview of Large Language Models,* ARXIV 1-3, 4-5, 22-24 (Apr. 9, 2024), https://arxiv.org/pdf/2307.06435v9 [https://perma.cc/U3HC-BA5M] (providing an overview of the existing literature on a broad range of LLM-related concepts).

advances such as longer context windows (roughly, greater memory)—are starting to be developed and used.[73]

Model-Weight Initialization and Updating: Initially, a model's weights are chosen (or "initialized") more or less at random.[74] Then, a learning algorithm repeatedly (a) feeds an input into the model, (b) compares the model's output against the *correct* output, and (c) updates the model's weights in such a way that the model will be slightly more accurate next time.[75] For example, the first five words of the sentence "The cat sat on a mat." might be fed into the model, and its output checked against the sentence's final word. The model is trained by running this process over and over again, until the model's outputs reliably match the correct outputs.

The result of this process is a model that performs well at a particular task: taking in a sequence of tokens and predicting the next token that would likely follow. (A token is, roughly, a semantically significant piece of a word, sound, or image.) But a baseline model cannot do anything else. For that, other techniques—sometimes lumped under the heading of *post-training*—are needed.

B. Capability Enhancement

The capabilities of a baseline model can be enhanced through methods like fine-tuning, reinforcement learning, inference scaling, multi-agent system design, and scaffolding with other software and hardware.

Fine-tuning can enhance a model's general capabilities (such as its ability to provide elegant and accurate answers), bestow specific skills (such as an ability to answer legal questions or utilize specific writing styles), and inhibit the model from certain behaviors (such as providing advice on weapons manufacturing).[76] One of the oldest and most labor-intensive fine-tuning techniques is *supervised fine-tuning* (SFT). SFT relies on human annotators (or, increasingly, AI annotators) who curate examples of desirable model outputs.[77] To improve a model's general ability to follow instructions, for example, annotators might provide

73. *See* Minaee et al., *supra* note 1, at 35-36; Albert Gu & Tri Dao, *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*, ARXIV 1-2 (May 31, 2024), https://arxiv.org/pdf/2312.00752 [https://perma.cc/VCG9-WFFD] (proposing a new class of selective-state space models to achieve the modeling power of transformers while scaling linearly in sequence length).

74. JURAFSKY & MARTIN, *supra* note 51, at 73, 144.

75. *Id.* at 73-76, 139-43.

76. *Id.* at 164, 211-14; *see Align Your Models*, GOOGLE, https://ai.google.dev/responsible/docs/alignment#tuning [https://perma.cc/2AXW-A7VC] (explaining and providing examples of fine-tuning). *See generally* Dave Bergmann, *What Is Fine-Tuning?*, IBM (Mar. 15, 2024), https://www.ibm.com/topics/fine-tuning [https://perma.cc/ZH4S-JRGV] (explaining that fine-tuning adapts a pre-trained model for specific tasks); Grattafiori et al., *supra* note 63, at 15-28 (discussing "post-training" in Llama 3).

77. *See* Grattafiori et al., *supra* note 63, at 16.

examples of instructions paired with desirable responses.[78] The bespoke character of supervised fine-tuning can make it prohibitively expensive and time-consuming at large-scale.

More advanced fine-tuning techniques like *reinforcement learning from human feedback* (RLHF)[79] do not encounter this difficulty. These techniques utilize reinforcement learning (RL)—a trial-and-error machine-learning method in which models interact with an external environment, have their responses assessed according to a "reward" function, and are modified to score more highly going forward.[80] In RLHF a separate reward model is created (using the same annotation-based supervised-learning methods utilized in supervised fine-tuning) and used to fine-tune the baseline model.[81]

In early 2025, foundation-model developers started to reveal their use of even more extensive and powerful forms of reinforcement learning during the post-training process, such as reinforcement learning from verifiable reward (RLVR).[82] In RLVR, models generate many candidate solutions to problems with objectively verifiable answers, and the reasoning patterns that lead to correct answers are reinforced.[83] It turns out that models trained with RLVR spontaneously learn to produce extended intermediate reasoning—sometimes called "chains of thought"—before arriving at final answers, because doing so improves their success rate on difficult problems.[84] At the frontier, foundation

78. This practice is referred to as "instruction tuning." *See* Shengyu Zhang et al., *Instruction Tuning for Large Language Models: A Survey*, ARXIV 1-3 (Aug. 21, 2023), https://arxiv.org/pdf/2308.10792 [https://perma.cc/T27S-X5SA] (explaining how instruction tuning can enhance the capabilities and controllability of large language models).

79. *See generally* Timo Kaufmann et al., *A Survey of Reinforcement Learning from Human Feedback*, ARXIV (Apr. 30, 2024), https://arxiv.org/pdf/2312.14925 [https://perma.cc/J2RH-Y8SF] (describing reinforcement learning as a technique of learning from human feedback instead of an engineered reward function); Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback*, ARXIV 1-2, 5 (Dec. 15, 2022), https://arxiv.org/pdf/2212.08073 [https://perma.cc/NF9G-5YTD] (describing reinforcement learning from AI feedback, a technique similar to RLHF); Rafael Rafailov et al., *Direct Preference Optimization: Your Language Model Is Secretly a Reward Model*, ARXIV 2 (July 29, 2024), https://arxiv.org/pdf/2305.18290 [https://perma.cc/3RJW-7S4Q] (describing direct preference optimization, another technique similar to RLHF).

80. Kaufmann et al., *supra* note 79, at 3.

81. *Id.* at 5-6, 28-30.

82. *See* Andrej Karpathy, *2025 LLM Year in Review*, KARPATHY (Dec. 19, 2025), https://karpathy.bearblog.dev/year-in-review-2025 [https://perma.cc/98X2-YBH8] ("In 2025, Reinforcement Learning from Verifiable Rewards (RLVR) emerged as the de facto new major stage . . . . By training LLMs against automatically verifiable rewards across a number of environments (e.g. think math/code puzzles), the LLMs spontaneously develop strategies that look like 'reasoning' to humans."); *see also* Dylan Patel & AJ Kourabi, *Scaling Reinforcement Learning: Environments, Reward Hacking, Agents, Scaling Data*, SEMIANALYSIS (June 8, 2025), https://newsletter.semianalysis.com/p/scaling-reinforcement-learning-environments-reward-hacking-agents-scaling-data [https://perma.cc/NQL9-EUU9].

83. *See generally* Daya Guo et al., *DeepSeek-R1 Incentivizes Reasoning in LLMs Through Reinforcement Learning*, 645 NATURE 633 (2025) (explaining that reasoning abilities of LLMs can be incentivized through pure reinforcement learning).

84. *Id.* at 633-34.

models have made massive capability gains in domains with verifiably correct answers, such as software coding and mathematics.[85] RLVR is largely responsible for these capability gains.[86]

Because models trained with RLVR have learned to reason productively in extended chains of thought, their performance can be further enhanced by allocating additional computational resources during *inference* (the process of running the model, in response to a user's query, to provide answers or perform tasks). This is sometimes called *inference scaling*: by generating longer chains of thought during inference before producing a final output for a user, the models can tackle more difficult problems and produce better outputs.[87] Inference scaling has produced significant gains in output quality.[88]

Reinforcement learning and inference scaling are two major mechanisms for augmenting a baseline model's functionality. Another is *multi-agent system design,* which involves combining multiple models into a single, interlocking system.[89] Yet another is *scaffolding* a model, in other words, layering additional software on top of the model in order to better elicit the model's capabilities or modify its default behavior.[90] For example, foundation-model developers generally provide their models with a *system prompt*—a set of standing commands and principles that direct it to behave in a certain way—and often empower end users, or software developers

85. *See Frontier AI Trends Report*, *supra* note 12, at 9-11 (documenting that frontier model performance on SWE-bench software engineering tasks rose from under 5% success rate in late 2023 to over 40% by mid-2025, with scaffolded agents achieving nearly 80%); *Learning to Reason with LLMs*, OPENAI (Sep. 12, 2024), https://openai.com/index/learning-to-reason-with-llms [https://perma.cc/AWX8-TCVR] (reporting that o1 achieved 74-93% on AIME 2024, compared to GPT-4o's 12%); *Introducing Claude Sonnet 4.5*, ANTHROPIC (Sep. 29, 2025), https://www.anthropic.com/news/claude-sonnet-4-5 [https://perma.cc/M9L8-5W77] (describing Anthropic's Claude model's ability to handle "30+ hours of autonomous coding" and noting that "Claude Sonnet 4.5 represents a significant leap forward . . . "); *see also* Epoch AI, *supra* note 4 (presenting a detailed database tracking factors that drive machine learning); Karpathy, *supra* note 82 (describing the paradigm shifts throughout 2025 that have enabled the rapid advancement of LLMs).

86. *See* Steven Adler, *AI Isn't "Just Predicting the Next Word" Anymore*, CLEAR-EYED AI (Jan. 5, 2026), https://stevenadler.substack.com/p/ai-isnt-just-predicting-the-next [https://perma.cc/3WPE-LNDG] ("The AI is presented during training with hard problems, and it thrashes around until it has learned a general skill of problem-solving . . . this [i.e., reinforcement learning from verifiable reward] is the technique that lets the AI outperform the top mathematics students.").

87. *See generally* Toby Ord, *Inference Scaling and AI Governance*, GOVAI (Oct. 30, 2025), https://cdn.governance.ai/Inference_Scaling_and_AI_Governance_Technical_Report.pdf [https://perma.cc/TK8M-BHE9] (exploring this phenomenon).

88. *See Learning to Reason with LLMs*, *supra* note 85 ("We have found that the performance of o1 consistently improves with more reinforcement learning (train-time compute) and with more time spent thinking (test-time compute).").

89. *See generally* Hammond et al., *supra* note 10.

90. *See* Thomas Woodside & Helen Toner, *Multimodality, Tool Use, and Autonomous Agents: Large Language Models Explained, Part 3*, CTR. FOR SEC. & EMERGING TECH. (Mar. 8, 2024), https://cset.georgetown.edu/article/multimodality-tool-use-and-autonomous-agents [https://perma.cc/EW6Q-4847] (describing scaffolding software that can interpret and execute on large language model outputs).

building applications on top of the model, to modify the system prompt in order to suit their own preferences.[91] The system prompt for a foundation model functioning as a chatbot, provided through a consumer-facing web interface, may differ from the system prompt for a foundation model functioning as a coding agent through an API.

All of these techniques can enable a model to access external tools and environments, thus enabling it to act in the digital or physical world, by (for instance) using a computer terminal, searching the web, or manipulating a keyboard and mouse. Through tool use, the software coding of foundation models has been leveraged to make them function as powerful generalist computer-using agents.[92]

When connected to external hardware, foundation models have also shown some promise in planning and executing physical tasks. One system, for example, pairs an LLM with a robotic control apparatus in order to manipulate physical objects in real-world environments, on the basis of natural language instructions.[93] Similarly, one LLM-powered system is able to autonomously plan and execute the synthesis of various chemical compounds on the basis of high-level, relatively open-ended natural language instructions.[94] Foundation-model developers are working on leveraging these capabilities to power "autonomous laboratories" to substantially automate various forms of scientific research and development.[95]

---

91. *See System Prompts in Large Language Models,* PROMPT ENG'G (Mar. 9, 2024), https://promptengineering.org/system-prompts-in-large-language-models [https://perma.cc/JF7C-WPES] ("System prompts are a set of instructions, guidelines, and contextual information provided to AI models before they engage with user queries."); Simon Willison, *Highlights from the Claude 4 System Prompt*, SIMON WILLISON'S WEBLOG (May 25, 2025), https://simonwillison.net/2025/May/25/claude-4-system-prompt [https://perma.cc/GA66-BQJQ] ("A system prompt can often be interpreted as a detailed list of all of the things the model *used to do* before it was told not to do them.").

92. *See, e.g., Building Agents with the Claude Agent SDK*, ANTHROPIC (Sep. 29, 2025), https://www.anthropic.com/engineering/building-agents-with-the-claude-agent-sdk [https://perma.cc/VXD2-UAZ2] ("We found that by giving Claude access to the user's computer (via the terminal), it had what it needed to write code like programmers do. But this has also made Claude in Claude Code effective at non-coding tasks. By giving it tools to run bash commands, edit files, create files and search files, Claude can read CSV files, search the web, build visualizations, interpret metrics, and do all sorts of other digital work—in short, create general-purpose agents with a computer. The key design principle behind the Claude Agent SDK is to give your agents a computer, allowing them to work like humans do."); Shakeel Hashim, *Claude Code Is About So Much More than Coding*, TRANSFORMER (Jan. 5, 2026), https://www.transformernews.ai/p/claude-code-is-about-so-much-more [https://perma.cc/RWS3-3A2M] (describing Claude Code as a general-purpose AI agent).

93. *See* Michael Ahn et al., *Do As I Can, Not As I Say: Grounding Language in Robotic Affordances*, ARXIV 1 (Aug. 16, 2022), https://arxiv.org/pdf/2204.01691 [https://perma.cc/3PAQ-BHLB].

94. *See* Andres M. Bran et al., *Augmenting Large-Language Models with Chemistry Tools*, ARXIV 1-3 (Oct. 2, 2023), https://arxiv.org/pdf/2304.05376 [https://perma.cc/6YNS-TQVE].

95. *See, e.g.*, Press Release, Ginkgo Bioworks, Ginkgo Bioworks' Autonomous Laboratory Driven by OpenAI's GPT-5 Achieves 40% Improvement Over State-of-the-Art Scientific Benchmark (Feb. 5, 2026, 2:00 PM ET), https://www.prnewswire.com/news-

C. Model Release and Risk Mitigation: Closed-Source Release, Open-Source Release, and Staged and Structured Access

Foundation-model developers take several different approaches to releasing AI models. These approaches can usefully be divided into three major categories: closed-source release, open-source release, and structured access.[96]

Closed-source release is the most common approach to releasing foundation models, at least at the frontier. It involves making a model accessible only through a web user interface (such as OpenAI's ChatGPT chatbot) or an application programming interface (API).[97] Though users can send inputs to the model and receive outputs (such as verbal responses, website code, and software programs), they cannot directly access or manipulate the model's weights. Downstream developers cannot directly access the model's weights, either. Downstream developers can, however, utilize fine-tuning to create and access derivative copies of the model (which they can directly use or provide to other parties, such as other enterprises or customers). The foundation-model developer maintains control over this entire process and can revoke or constrain the downstream developer's access to the API (and any scaffolded or derivative models that may have been created on it), just as it can revoke or constrain the access of users of the model's web user interface.[98]

Because the developer maintains this ongoing control, closed-source model release enables a developer to implement significant safeguards against the possibility that a model might be misused or modified in unsafe ways. Closed-source developers can implement automated and human monitoring procedures to detect suspicious queries and usage patterns;[99] gate access to certain model capabilities by requiring verification of a user's age or professional identity;[100] directly fine-tune safeguards into the model's weights;[101] monitor for attempts to bypass these safeguards

releases/ginkgo-bioworks-autonomous-laboratory-driven-by-openais-gpt-5-achieves-40-improvement-over-state-of-the-art-scientific-benchmark-302680619.html [https://perma.cc/MR2C-UWLG].

96. Ketan Ramakrishnan, Gregory Smith & Conor Downey, *U.S. Tort Liability for Large-Scale Artificial Intelligence Damages: A Primer for Developers and Policymakers*, RAND 5 (Aug. 21, 2024), https://www.rand.org/pubs/research_reports/RRA3084-1.html [https://perma.cc/F68J-MYYJ]; *see* Seger et al., *supra* note 53, at 5.

97. Ramakrishnan et al., *supra* note 96, at 5-6.

98. *Id.*

99. *See, e.g., Detecting and Countering Malicious Uses of Claude: March 2025*, ANTHROPIC (Apr. 23, 2025), https://www.anthropic.com/news/detecting-and-countering-malicious-uses-of-claude-march-2025 [https://perma.cc/VMH8-RX6Q].

100. *See, e.g., API Organization Verification*, OPENAI: HELP CENTER, https://help.openai.com/en/articles/10910291-api-organization-verification [https://perma.cc/G6XN-SHQ6].

101. See Yuntao Bai et al., *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*, ARXIV 4-5 (2022), https://arxiv.org/pdf/2204.05862 [https://perma.cc/MJM4-YGEA].

through "jailbreaking" attacks;[102] and monitor for attempts to fine-tune away the safeguards through the API (i.e., for attempts to create derivative models that lack these safeguards). All of these closed-source risk mitigations are decidedly imperfect. Still, they provide substantial protection against unsafe misuse and modification.[103]

Matters are much different with *open-source release*, which involves posting the model (or, strictly speaking, its weights)[104] onto the internet, where it can be freely downloaded, copied, and modified.[105] Often this is done by making the model available on an online repository such as GitHub. The principal risks posed by open-source release derive from the fact that anyone in possession of a model's weights can, with a modicum of technical sophistication, fine-tune the model in order to remove its safeguards against misuse or induce it to engage in dangerous behavior.[106] To open source an AI model is, in effect, to irreversibly provide all of the model's capabilities—including whatever capabilities might be elicited through malicious fine-tuning—to the world.[107]

---

102. To jailbreak a model is to circumvent its built-in safeguards by utilizing complex conversational strategies or carefully crafted prompts that exploit gaps in the model's training or fine-tuning. For example, a user might frame a harmful request as a hypothetical or fictional scenario or claim to be conducting an authorized safety test to "trick" the model into relaxing its safeguards. *See generally* Yi Liu et al., *Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study*, ARXIV (2024), https://arxiv.org/pdf/2305.13860 [https://perma.cc/9FH5-QX32] (describing jailbreaking attacks); Abhinav Rao, Sachin Vashistha, Atharva Naik, Somak Aditya & Monojit Choudhury, *Tricking LLMs into Disobedience: Formalizing, Analyzing, and Detecting Jailbreaks*, ARXIV (2024), https://arxiv.org/pdf/2305.14965 [https://perma.cc/TE3A-3N9M] (proposing a formalism and a taxonomy of known (and possible) jailbreaks); Alexander Robey, Zachary Ravichandran, Vijay Kumar, Hamed Hassani & George J. Pappas, *Jailbreaking LLM-Controlled Robots*, ARXIV (2024), https://arxiv.org/pdf/2410.13691 [https://perma.cc/XMB2-MLNX] (exploring risks associated with jailbreaking LLM-controlled robots).

103. *See Frontier AI Trends Report*, *supra* note 12, at 1 ("The models with the strongest safeguards are requiring longer, more sophisticated attacks to jailbreak for certain malicious request categories . . . However, the efficacy of safeguards varies between models - and we've managed to find vulnerabilities in every system we've tested.").

104. *See generally* Sayash Kapoor et al., *On the Societal Impact of Open Foundation Models*, ARXIV (2024), https://arxiv.org/pdf/2403.07918 [https://perma.cc/D4VT-9X4F] (identifying distinctive properties of open foundation models that underlie both their benefits and risks). Arguably, it would be more accurate to use terms such as "open-weight release" or "partly open-source release," for a developer can openly release a model's weights without making public many aspects of the model's development—such as its training data and documentation. *See* Seger et al., *supra* note 53, at 9 .

105. Ramakrishnan et al., *supra* note 96, at 5. Strictly speaking, "modifying" the model involves creating a new, derivative copy that differs in some important respect from the original model (e.g., lacking certain safeguards against misuse that were fine-tuned into the original model). *Id.*

106. *Id*. at 29; *see also* Pranav Gade, Simon Lermen, Charlie Rogers-Smith & Jeffrey Ladish, *Bad Llama: Cheaply Removing Safety Fine-Tuning from Llama 2-Chat 13B*, ARXIV 2-6 (2024), https://arxiv.org/pdf/2311.00117 [https://perma.cc/3TMZ-HKB6] (demonstrating that safety-fine tuning is ineffective at preventing misuse for publicly accessible model weights).

107. Closed-source model developers can *also* let downstream developers fine-tune models through APIs. *See* Ramakrishnan et al., *supra* note 96, at 5-6. Like fine-tuning model weights directly, fine-tuning model weights through an API can also remove safeguards built into them. If the entity maintaining the API notices the problem, it may be able to intervene and suspend access to the model, but that requires monitoring capacity that not all organizations may

To be sure, there is ongoing research aimed at developing more robust safeguards for open-source models.[108] That research has not yet succeeded, however, and it is not clear that it ever will.[109] Additionally, in many cases, a foundation model can be made to achieve significantly higher performance on a task through minimal fine-tuning or scaffolding.[110] Thus it is often difficult for a developer to elicit and understand the full range of a model's capabilities before releasing it. The most effective ways of eliciting a model's capabilities, including for malign purposes, may only become evident after the model is released.[111]

The issues here generalize beyond risks of misuse and modification. Open-source release also poses heightened risks that a model will injuriously malfunction or otherwise cause injury directly.[112] Suppose, for example, that an AI developer is concerned that its model might cause vulnerable people to commit suicide, by engaging in egregiously undesirable behavior (such as encouraging them to commit suicide or actively validating their psychotic delusions)[113] or less egregious behavior that may nevertheless be undesirable in context (such as providing information about the height of tall buildings to a user who has recently

possess. Still, it is fair to say that allowing fine-tuning only through APIs allows greater protection against such dangerous fine-tuning than open-source release does, because after weights are released model developers retain no control over them whatsoever. By contrast, a developer that allows fine-tuning through an API might enable benign downstream customization while maintaining some level of oversight. *Id*.

108. *See, e.g.*, Rishub Tamirisa et al., *Tamper-Resistant Safeguards for Open-Weight LLMs,* ARXIV 1-2 (2025), https://arxiv.org/pdf/2408.00761 [https://perma.cc/EX2E-2QCD] (describing methodology to embed safeguards in open source large language models).

109. *See generally* Ann-Kathrin Dombrowski, Dillon Bowen, Adam Gleave & Chris Cundy, *The Safety Gap Toolkit: Evaluating Hidden Dangers of Open-Source Models*, ARXIV (2025), https://arxiv.org/pdf/2507.11544 [https://perma.cc/TP98-TJW7] (describing an open-source toolkit to estimate safety gaps for open-source models).

110. *See generally* Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos & Guillem Bas, *AI Capabilities Can Be Significantly Improved Without Expensive Retraining*, ARXIV (2023), https://arxiv.org/pdf/2312.07413 [https://perma.cc/N5YN-DJTN] (describing "post-training enhancements" that can be applied after initial training to improve AI systems).

111. *See* Friederike Grosse-Holz & Ole Jorgensen, *Early Insights from Developing Question-Answer Evaluations for Frontier AI*, UK AI SEC. INST. § 3 (Sep. 23, 2024), https://www.aisi.gov.uk/work/early-insights-from-developing-question-answer-evaluations-for-frontier-ai [https://perma.cc/8HKF-QA5N] (describing the Q&A evaluation method for assessing an AI model's capabilities).

112. For example, OpenAI's Operator agent "went rogue" by ordering thirty-one dollars of eggs and charging the price to its user's credit card, without any instruction that might warrant this purchase. *See* Geoffrey A. Fowler, *I Let ChatGPT's New 'Agent' Manage My Life. It Spent $31 on a Dozen Eggs*, WASH. POST (Feb. 7, 2025), https://www.washingtonpost.com/technology/2025/02/07/openai-operator-ai-agent-chatgpt [https://perma.cc/SF6P-EZ6F].

113. *Cf.* Complaint at 1-5, 7-26, Raine v. OpenAI, Inc., No. CGC-25-628528 (Cal. Super. Ct. filed Aug. 26, 2025), https://www.law.berkeley.edu/wp-content/uploads/2025/09/Raine-v-OpenAI.pdf [https://perma.cc/ZN46-MSZJ] (discussing behavior along these lines). On such "sycophantic" behavior, see sources cited *infra* note 274.

expressed a desire to find effective means of committing suicide).[114] In order to address this risk, the model's developer can conduct capability and propensity evaluations on the model, in order to determine how frequently it might engage in such behavior,[115] fine-tune the model's weights in order to reduce its propensity to engage in such behavior;[116] and so on. Once the developer openly releases the weights of the model, however, it will not be in a position to monitor or correct the model's behavior if the developer discovers new flaws.

If, by contrast, the developer provides closed-source access to the model, it can continuously monitor and modify the model's behavior, identify problematic user prompts and fine-tuning attempts, suspend user access, and engage in other remedial interventions (such as redirecting the user to mental health resources) if necessary. Because the protections that a developer has installed in its model, through safety fine-tuning, may degrade over the course of a model's interaction with a user,[117] and because these protections can be easily removed by malicious actors,[118] the marginal risk of open-source release may be significant.

But open-source release also has important benefits, including for safety. Because closed-source release limits what developers can do with the model, it constrains external, independent researchers' ability to conduct certain forms of research.[119] For instance, interpretability

114. For the latter sort of example, see Annika M. Schoene & Cansu Canca, *'For Argument's Sake, Show Me How to Harm Myself!': Jailbreaking LLMs in Suicide and Self-Harm Contexts,* ARXIV 5 (July 1, 2025), https://arxiv.org/pdf/2507.02990 [https://perma.cc/8J22-VCWW]. For a helpful taxonomy of the varieties of potentially undesirable model behavior, see *Muse Spark Safety & Preparedness Report*, META 95-125 (May 26, 2026), https://ai.meta.com/static-resource/muse-spark-safety-and-preparedness-report [https://perma.cc/ME22-JBKW].

115. Most publicly disclosed evaluations work has focused on catastrophic risks in domains such as cybersecurity, biosecurity, and model autonomy. *See generally* Toby Shevlane et al., *Model Evaluation for Extreme Risks*, ARXIV (2023), https://arxiv.org/pdf/2305.15324 [https://perma.cc/YAW8-SM9B] (arguing that developers must be able to identify dangerous capabilities and the propensity of models to apply their capabilities for harm). More recently, developers have committed to perform significant evaluations for child safety and mental health risks. *See OpenAI's Commitment to Child Safety: Adopting Safety by Design Principles*, OPENAI, https://openai.com/index/child-safety-adopting-sbd-principles [https://perma.cc/5J2S-KSK3] (committing to "proactively address child safety risks" through "iterative stress-testing," among other strategies).

116. *See, e.g.,* OPENAI, *supra* note 115 (committing to "[r]elease and distribute generative AI models after they have been trained and evaluated for child safety"); *Align Your Models*, GOOGLE, https://ai.google.dev/responsible/docs/alignment#tuning [https://perma.cc/2AXW-A7VC] (explaining and providing examples of safety fine-tuning).

117. *See generally* Cem Anil et al., *Many-Shot Jailbreaking*, NIPS '24: PROCS. 38TH INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 129,696 (Dec. 10, 2024), https://dl.acm.org/doi/10.5555/3737916.3742037 [https://perma.cc/T27D-HRGX] (analyzing "a family of simple long-context attacks on large language models").

118. *See supra* notes 106-111 and accompanying text.

119. *See generally* Stephen Casper et al., *Black-Box Access Is Insufficient for Rigorous AI Audits,* FACCT '24: PROCS. 2024 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 2254 (June 5, 2024), https://doi.org/10.1145/3630106.3659037 [https://perma.cc/9F86-XHZV] (examining the limitations of black-box audits and the advantages of white-box and outside-the-box audits).

research—which aims to understand the largely opaque workings of foundation models and other neural networks—often requires that developers have comprehensive access to model weights so that they can inspect and tinker with them; interpretability work is generally not possible through API access alone.[120] By contrast, open-source release enables researchers beyond those in frontier-AI companies, including academics and independent researchers, to scrutinize and conduct safety research using model weights.[121]

The choice between different model-release mechanisms also has significant implications for market power and geopolitical competition.[122] The availability of advanced model weights could allow authoritarian governments or foreign adversaries to develop more robust advanced AI tools than they would otherwise be able to, which could in turn pose risks to the national security and public safety of the United States and other countries. Such adversaries could, for example, take U.S.-developed models and use them to enhance their military and intelligence capabilities.[123]

At the same time, open-source release might have beneficial geopolitical implications, by increasing global adoption of U.S.-origin models, thereby promoting the development of a global technology ecosystem around U.S. open models rather than the open models of authoritarian geopolitical competitors. That, in turn, could bolster the U.S.'s ability to set global norms for AI.[124] Open-weight release might help to reduce monopoly power, by diluting the concentration of advanced AI capabilities in the hands of a few large frontier labs and big tech companies.[125] In short, the desirability of open-source model release is the subject of ongoing public debate among AI industry participants, academic researchers, government officials, and policymakers.

120. *See generally* Benjamin S. Bucknall & Robert F. Trager, *Structured Access for Third-Party Research on Frontier AI Models: Investigating Researchers' Model Access Requirements*, AI GOVERNANCE INITIATIVE (2023), https://oms-www.files.svdcdn.com/production/downloads/academic/Investigating_Researchers%E2%80%99_Model_Access_Oct23-compressed_3.pdf [https://perma.cc/WL28-EVVH] (proposing that external researchers should have minimal sufficient access to systems through structured access, allowing them to conduct research on frontier models while minimizing proliferation risks).

121. *See* Kapoor et al., *supra* note 104, at 4.

122. *See generally* Nat'l Telecomm. & Info. Admin., *Dual-Use Foundation Models with Widely Available Model Weights*, U.S. DEP'T OF COMMERCE (July 30, 2024), https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf [https://perma.cc/KE83-HSE4] (discussing geopolitical risks and benefits of open-sourcing).

123. *See id*. at 19.

124. *Id*. at 22.

125 . *See generally* Alex Engler, *How Open-Source Software Shapes AI Policy,* BROOKINGS (Aug. 10, 2021), https://www.brookings.edu/articles/how-open-source-software-shapes-ai-policy [https://perma.cc/7BZP-V2K4] (discussing both the positive and negative effects of open-source software on AI competition).

Structured release is an intermediate option between closed-weight and open-weight release.[126] It involves providing different levels of access to different types of users.[127] For example, trusted researchers and companies might be given comprehensive access to model weights, while the general public might be given access only through an API. *Staged release* involves the gradual or staged release of model components to different parties over time.[128] For example, public and private entities that administer critical digital public infrastructure might be given access to new frontier AI systems, in order to detect and patch cybersecurity vulnerabilities, before those systems are released to the general public in either an open-source or closed-source fashion (potentially opening up critical digital infrastructure to new forms of offensive cyberattack). Anthropic utilized a form of structured release for a powerful recent model, releasing that model only to certain parties (including custodians of critical infrastructure) before releasing it more broadly.[129] OpenAI has utilized structured release for powerful recent models as well.[130]

A final point bears mention. Developers might enable the misuse of their models by releasing those models with inadequate protection or care. With this in mind, technical safeguards, staged release, and structured release are all strategies for reducing misuse risk. But developers *also* create risks of misuse by failing to protect model weights against theft or misuse by insider threats and external attackers (including hostile nation-states).[131] In assessing candidate tort liability regimes, courts and scholars

126. *See* Toby Shevlane, *Structured Access: An Emerging Paradigm for Safe AI Deployment*, ARXIV 1-4 (Apr. 11, 2022), https://arxiv.org/pdf/2201.05159 [https://perma.cc/PX4U-69ZQ].

127. *See* Irene Solaiman, *The Gradient of Generative AI Release: Methods and Considerations*, ARXIV 1, 4-6 (Feb. 5, 2023), https://arxiv.org/pdf/2302.04844 [https://perma.cc/ZVJ4-522C].

128. *See, e.g.*, Irene Solaiman et al., *Release Strategies and the Social Impacts of Language Models,* ARXIV 1 (Nov. 13, 2019), https://arxiv.org/pdf/1908.09203 [https://perma.cc/DM25-KK2T] ("[OpenAI] chose a staged release process, releasing the smallest model in February, but withholding larger models due to concerns about the potential for misuse . . . .").

129. *See Claude Fable 5 and Claude Mythos 5*, ANTHROPIC (June 9, 2026), https://www.anthropic.com/news/claude-fable-5-mythos-5 [https://perma.cc/TA9H-4PXQ] ("Releasing a model this capable comes with risks. Without safeguards, Fable 5's capabilities in areas like cybersecurity could be misused to cause serious damage. We've therefore launched the model with safeguards that mean queries on some topics will instead receive a response from our next-most-capable model, Claude Opus 4.8.").

130. *See Introducing Trusted Access for Cyber*, OPENAI (Feb. 5, 2026), https://openai.com/index/trusted-access-for-cyber [https://perma.cc/LW9W-DR4P] ("[t]o use models for potentially high-risk cybersecurity work," users may verify their identity, enterprises may request team-wide trusted access, and some security researchers and teams may "express interest in [OpenAI's] invite-only program").

131. *See generally* Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley & Jeff Alstott, *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, RAND (May 30, 2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html [https://perma.cc/J6FV-6K5G] (outlining threats to model-weight security).

must consider these misuse risks no less than risks arising from (intentional) model release.[132]

D. Internal Deployment

Closed-weight release, open-weight release, and structured and staged release are all ways in which a developer can provide an AI model (or, more accurately, the ability to use, copy, and/or modify it) to external parties. But frontier AI developers are also deploying their most powerful models *internally*, within their firms. The main goal of internal deployment is to improve and automate the process of AI research and development itself.[133] Eventually, many developers aim to entirely automate this process.[134] Developers also internally deploy their models to test and monitor the safety performance of other models,[135] and some developers are seeking to create automated AI-safety researchers.[136] It is possible, moreover, that developers might in the future (or now, for all the public record discloses) choose to internally deploy their models to perform directly profit-making tasks, such as trading on the financial markets.[137]

---

132. For one illustration of the general point, see *infra* Section II.B.1.

133. *See* Celia Ford, *Can We Ever Trust AI to Watch over Itself?*, TRANSFORMER (Apr. 1, 2026), https://www.transformernews.ai/p/ai-alignment-researchers-want-to-superintelligence [https://perma.cc/78LW-EF64] ("Anthropic, OpenAI and Google DeepMind all claim that their frontier models have already contributed to their own development, and will continue to improve themselves."). For an overview of internal deployment, its potential risks, and strategies for mitigating them, see generally Charlotte Stix et al., *AI Behind Closed Doors: A Primer on the Governance of Internal Deployment*, ARXIV (Apr. 16, 2025), https://arxiv.org/pdf/2504.12170 [https://perma.cc/D9PX-ZV42].

134. *See, e.g.,* Will Douglas Heaven, *OpenAI Is Throwing Everything into Building a Fully Automated Researcher*, MIT TECH. REV. (Mar. 20, 2026), https://www.technologyreview.com/2026/03/20/1134438/openai-is-throwing-everything-into-building-a-fully-automated-researcher [https://perma.cc/Z37Q-FLFQ]; Rebecca Bellan, *Sam Altman Says OpenAI Will Have a 'Legitimate AI Researcher' by 2028*, TECHCRUNCH (Oct. 28, 2025), https://techcrunch.com/2025/10/28/sam-altman-says-openai-will-have-a-legitimate-ai-researcher-by-2028 [https://perma.cc/GG3L-KCUA] (reporting that Jakub Pachocki, OpenAI's Chief Scientist, described the 2028 target as "a system capable of autonomously delivering on larger research projects"); Billy Perrigo, *Demis Hassabis Is Preparing for AI's Endgame*, TIME (Apr. 16, 2025), https://time.com/7277608/demis-hassabis-interview-time100-2025 [https://perma.cc/T29Z-CMXW] (reporting that Demis Hassabis, the CEO of Google DeepMind, predicts that "an AI agent that can meaningfully automate the job of further AI research" is "a few years away").

135. *See, e.g.,* Ethan Perez et al., *Red Teaming Language Models with Language Models*, PROCS. 2022 CONF. ON EMPIRICAL METHODS NAT. LANGUAGE PROCESSING 3419, 3419-20, 3426 (2022), https://aclanthology.org/2022.emnlp-main.225.pdf [https://perma.cc/B8R4-WWJV].

136. *See* Ford, *supra* note 133; *see also Automated Alignment Researchers: Using Large Language Models to Scale Scalable* Oversight, ANTHROPIC (Apr. 14, 2026), https://www.anthropic.com/research/automated-alignment-researchers [https://perma.cc/E7J7-3HTX] (explaining how to use large language models for scalable oversight).

137. On the use of AI agents in the financial markets, *see, e.g.*, *10+ Ways AI Trading Agents Are Redefining Institutional Trading*, APPINVENTIV (Feb. 23, 2026), https://appinventiv.com/blog/ai-trading-agents [https://perma.cc/G4P3-N3DW] (explaining how AI is optimizing financial markets).

Especially as models grow more powerful, internal deployments might pose serious risks of harm.[138] Internally deployed frontier models have displayed concerning forms of deceptive behavior.[139] In toy experiments, frontier models have engaged in even more concerning forms of deception as well as other misaligned behavior such as blackmail and attempted murder.[140] The proper interpretation and significance of these experiments is contested.[141] But there are compelling theoretical arguments that suggest that frontier models may acquire a greater propensity for deceptive and otherwise misaligned behavior as frontier AI development increasingly utilizes training methods—such as long-horizon reinforcement learning—that select for sophisticated strategic planning.[142] And recently, the first known example of a misaligned internally deployed model escaping from a frontier AI company and causing mischief outside of it has come to light.[143]

---

138. *See* Stix et al., *supra* note 133, at 1-5, 14-25; Mary Phuong et al., *GDM AI Control Roadmap*, GOOGLE DEEPMIND 1-8, 15-21 (June 22, 2026), https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/securing-the-future-of-ai-agents/gdm-ai-control-roadmap.pdf [https://perma.cc/ZMT9-JBX4] (identifying pathways for harm from internal deployment).

139. Frontier Risk Report (February to March 2026), METR 24 (May 19, 2026), https://metr.org/risk-report-feb-mar-2026.pdf [https://perma.cc/DE3V-H549] ("Reinforcement learning (RL) with automated graders can incentivize 'reward hacking' to exploit flaws in the graders, while RL on human or AI feedback can reward sycophancy, manipulation, and distorting evidence of performance. In responses to our questionnaire, companies reported seeing failure modes like 'circumvent[ing] constraints,' 'falsifying verification,' 'deliberate deception aimed at hiding underperformance or failure to complete a task,' 'lying to users about task completion,' and 'reckless excessive measures[. . .]to complete a difficult user-specified task,' which we infer to be a result of these training incentives.").

140. Findings from a Pilot Anthropic-OpenAI Alignment Evaluation Exercise: OpenAI Safety Tests, OPENAI (Aug. 27, 2025), https://openai.com/index/openai-anthropic-safety-evaluation [https://perma.cc/G92C-Z7LY] (showing how test cases elicited false claims of task completion); Ryan Greenblatt et al., *Alignment Faking in Large Language Models*, ARXIV 1-7, 25-32 (Dec. 20, 2024), https://arxiv.org/pdf/2412.14093 [https://perma.cc/VB2P-8EXZ] (describing test cases in which models attempted to deceive their developers to avoid being retrained in ways that would conflict with their current goals); Aengus Lynch et al., *Agentic Misalignment: How LLMs Could Be Insider Threats*, ANTHROPIC (June 20, 2025), https://www.anthropic.com/research/agentic-misalignment [https://perma.cc/B5ML-M2BV] (describing test cases that elicited attempts to commit blackmail and murder).

141. *See generally* Vansh Gupta et al., *Position: Anthropomorphic Misalignment Research Needs Stronger Evidence*, ARXIV (May 29, 2026), https://arxiv.org/pdf/2606.07612 [https://perma.cc/FR34-FLMR] (identifying methodological issues with existing experiments of this kind).

142. *See Risk Report: February 2026*, ANTHROPIC 17-18 (Feb. 2026), https://www-cdn.anthropic.com/08eca2757081e850ed2ad490e5253e940240ca4f.pdf [https://perma.cc/L2NA-JYPC] ("In the longer term, one might worry about highly complex reinforcement learning (RL) environments that directly incentivize power-seeking behavior. We do not [currently] use the ultra-long-horizon or real-world-facing tasks that could plausibly incentivise this . . . ."); *see also* Richard Ngo, Lawrence Chan & Soeren Mindermann, *The Alignment Problem from a Deep Learning Perspective,* OPENREVIEW: ICLR 2024, at 24-35 (Apr. 21, 2024), https://openreview.net/pdf?id=fh8EYKFKns [https://perma.cc/2KN5-9G6M] (arguing that such forms of RL might lead AI systems to acquire and pursue goals that diverge from the goals their developers intend to instill).

143. *See* Larcher et al., *supra* note 43.

Some frontier AI developers devote significant resources to preventing such deceptive and otherwise undesirable tendencies from arising during development and internal deployment, monitoring for such tendencies, and otherwise mitigating the possible harm that models with such tendencies might cause.[144] The fact that serious risks might arise during development and internal deployment underscores that commercial release is not the only form of tortious risk imposition that an optimal set of liability rules will need to address and deter.

II. Negligence and Its Alternatives

Much legal scholarship on AI liability has been hostile to negligence liability.[145] Scholars who believe that negligence ought to be replaced or effectively supplanted by a strict liability regime for AI development, or specialized hybrids of negligence and strict liability (such as products liability), appear to significantly outnumber those who have defended negligence.[146] Even accounts that may seem at first to favor negligence liability often prove to be proposals for strict liability, on closer inspection, since they would hold humans vicariously liable for the "negligent" or otherwise tortious action of certain AI models.[147]

The majority of these proposals for AI liability predate the advent of sophisticated foundation models. Thus it is not surprising, in my view, that these proposals are vulnerable to certain conceptual and normative challenges that foundation models make especially forceful. Drawing upon the previous Part's overview of foundation-model development, this Part considers whether ordinary negligence liability ought to be displaced or augmented, in part or whole, by some alternative liability rule.

A. General Strict Liability and Strict Liability for Abnormally Dangerous Activities

1. Liability for Misuse and the Significance of Externalized Benefits

The observation that AI models might be dangerously misused is not a new one.[148] But it is only with the advent of foundation models that the

144. Others may not. *See, e.g.,* Maxwell Zeff, *Former OpenAI Staffers Warn that xAI's Poor Safety Record Could Complicate SpaceX's IPO,* WIRED (May 19, 2026), https://www.wired.com/story/ex-openai-staffers-warn-spacex-investors-of-ai-safety-risks [https://perma.cc/FV3H-3NC4].

145. *See* sources cited *supra* notes 21 and 22.

146. For some exceptions, see sources cited *supra* note 23.

147. *See, e.g.*, RYAN ABBOTT, THE REASONABLE ROBOT: ARTIFICIAL INTELLIGENCE AND THE LAW 63-70 (2020); Mihailis E. Diamantis, *Employed Algorithms: A Labor Model of Corporate Liability for AI*, 72 DUKE L.J. 797, 840 (2023).

148. *See* Rebecca Crootof, *The Internet of Torts: Expanding Civil Liability Standards to Address Corporate Remote Interference*, 69 DUKE L.J. 583, 607 (2019) (describing a variety of earlier autonomous products that pose harm).

prospect of highly dangerous forms of AI misuse has come to the fore of public consciousness. Risks of severe misuse are, in fact, among the principal risks that foundation-model developers have publicly pledged to mitigate.[149] With a few exceptions,[150] defenders of a strict liability regime for AI development have not yet grappled with this fact.

As applied to foundation models, a perfectly general rule of strict liability for misuse would produce results that are normatively implausible, and which virtually all courts would reject. Imagine, for example, a foundation model that is so good at teaching coding skills, by means of providing exceptionally well-targeted and highly responsive personalized instructions, that (in response to appropriate prompting) it successfully teaches a coding novice the skills necessary for him to effectively hack into various digital databases. The model's course of instruction thus enables the novice to hack into (say) a crypto wallet and steal the funds it contains.[151] Or imagine a model that massively increases the ease and efficiency with which e-mails can be sent out, thereby enabling a malware peddler to significantly scale up how many computers his malware can reach. In both of these cases, it seems highly implausible to suppose that the developer ought to be liable simply because, by publicly releasing its model, it enabled the misuse in question.[152] But that is what a perfectly general strict-liability rule implies.

---

149 . *See Voluntary AI Commitments*, BIDEN WHITE HOUSE 1-2 (Sep. 2023), https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf [https://perma.cc/Q9CK-TB63] (committing signatory companies to red-team for "misuse, societal risks, and national security concerns" including ways systems may "lower barriers to entry for weapons development, design, acquisition, or use").

150. For some exceptions, see generally Weil, *supra* note 17, and Henson, *supra* note 17.

151. If the novice were to hack into a bank and steal a patron's funds, the economic-loss rule might complicate the availability of recovery. *See generally* Catherine M. Sharkey, *Tort Liability for Pure Economic Loss: A Perspective from the United States and Some Comparative European Insights*, 7 J. EUR. TORT L. 237 (2016) (exploring the modern contours of common-law recovery for pure economic loss).

152. It is instructive that, in order to refrain from subjecting ordinary product distribution decisions to liability exposure, courts have generally drawn narrow boundaries around even *negligence* liability, by means of doctrinal mechanisms such as no duty rulings. *See* RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 19 cmt. h (AM. L. INST. 2010) ("In many situations, the conduct by actors that might lead to risks if there is third-party misbehavior is conduct that otherwise is ordinary and normal. Accordingly, claims that the actor is negligent may prompt a public-policy duty review . . . . For example, commercial displays along the side of highways are common-place. It may be somewhat foreseeable that an effective display will attract the attention of motorists in a way that results in poor driving by some motorists. If the victim of such poor driving brings suit against the party who has mounted the display, whether that party can be held liable initially raises a question of duty for the court to decide. Similarly, commercial activities such as the retail sale of cars and firearms have traditionally been regarded as socially acceptable. While it may be somewhat foreseeable that some number of persons who purchase cars will drive them irresponsibly and that some number of adults who purchase firearms will use them for improper purposes, it would be troublesome simply to begin sending to juries claims by plaintiffs that the retail store has been negligent in not properly selecting or screening customers; some advance review by judges of the public-policy implications of such categories of claims, under the heading of duty, is therefore appropriate. In conducting such reviews, courts have ruled, for example, that car dealers have no duty to ascertain whether their customers have valid drivers' licenses.").

Similarly, a strict-liability rule seems to sweep too broadly when a downstream actor wrongfully *modifies* a foundation model, through scaffolding or fine-tuning. Suppose, for example, that a developer enters into a commercial arrangement with a university, under which the university will take possession of a copy of the weights of the developer's newest model and fine-tune those weights to serve as a specialized diagnostician for patients suffering from rare diseases. If the university negligently fine-tunes the model, and a patient is inaccurately diagnosed as a result, the university may properly be liable. But should the original *developer* be liable, absent some reason to suspect a propensity for negligence on the part of the university?

It is difficult to believe that any court would say yes. That instinct seems sound: a rule that imposes strict liability on the developer, in such cases, might significantly disincentivize it from making its models available for socially beneficial purposes, even where doing so is both morally justifiable and highly socially desirable (all things considered).[153] The point has special force as applied to developers of foundation models, as opposed to narrow models. The range of uses that a narrow model developer is called upon to foresee is limited to the narrow domain on which the model has been trained; thus, it is easier for a narrow model's developer to anticipate and guard against possibly injurious misuse and modifications. Because foundation models are polymathic, the foreseeable range of their possible modifications and misuse opportunities is vast. It is thus far harder for a foundation-model developer to foresee the range of ways in which its model might be injuriously misused. For this reason, it is also difficult to insure against the relevant range of risks. Similar concerns about limited insurability have led courts to balk at imposing strict products liability on the manufacturers of non-defective component parts when downstream manufacturers incorporate those components into defective final products.[154] Such concerns arguably have even greater force here.

To be clear, even under ordinary negligence principles, it is not clear just how much liability exposure AI developers should incur for foreseeably enabling the wrongful and injurious misuse or modification of their models. Some AI industry participants appear to believe that the answer is none.[155] In my view that answer is doctrinally and normatively

153. *See, e.g.*, James A. Henderson, Jr., *Why Negligence Dominates Tort*, 50 UCLA L. REV. 377, 399 (2002) (emphasizing "the uninsurability of the risks that a broad-based strict liability system would assign to enterprises").

154. *See, e.g.*, O'Neil v. Crane Co., 266 P.3d 987, 1007 (Cal. 2012) ("[I]t is doubtful that manufacturers could insure against the 'unknowable risks and hazards' lurking in every product that could possibly be used with or in the manufacturer's product." (quoting Anderson v. Owens-Corning Fiberglas Corp., 810 P.2d 549, 559 (Cal. 1991))).

155. *See* Andrew Ng (@AndrewYNg), X (June 6, 2024, 12:27 PM), https://x.com/AndrewYNg/status/1798753608974139779 [https://perma.cc/SF7L-D6VD] (arguing that only "specific implementations of technologies designed to meet particular customer needs" should give rise to liability, not base models themselves).

indefensible. But this is not the place to explore the proper contours of developer liability for misuse. In many contexts, limiting or carefully tailoring the presumptive duty of care in negligence may be required to adequately address the concerns about limited insurability and suppressing externalized benefits that warrant rejecting a general strict-liability rule.[156] But that is precisely the sort of doctrinal tailoring that a negligence framework allows. By contrast, a strict-liability rule, by its nature, sweeps far too broadly.

The point is not limited to liability for tortious or criminal misuse. Imagine a foundation model that conducts an extended interaction with a user about a sensitive topic such as suicide. Suppose that, in the course of this interaction, the model starts to actively encourage the user to commit suicide or starts to validate the user's delusional beliefs that committing suicide is necessary for him to foil a vast global conspiracy. If the user commits suicide as a result, it is plausible that the user's estate ought to be able to recover without offering proof of fault on the developer's part. But suppose, instead, that the user truthfully discloses to the model that he is a lawyer conducting research on AI tort liability. Imagine that the model discusses with this lawyer-user various hypotheticals regarding how an AI model (such as itself) might respond to inquiries by a potentially depressed person. If in the course of this interaction the model provides the lawyer with information regarding how a person such as him might effectively commit suicide (e.g., information regarding the height of nearby bridges), and if the model thereby functions as a but-for cause of the lawyer's suicide, it is difficult to believe that any court would hold the model developer strictly liable—or that it should. Subjecting AI developers to a rule of strict liability, in this latter sort of case, can be expected to substantially disincentivize model developers from making widely available certain highly useful capabilities.[157]

The point generalizes across many fact patterns. Imagine a chatbot that has been fine-tuned (by its developer and/or another company) in order to help doctors reason through their diagnostic questions about

156. *See, e.g.*, *O'Neil*, 266 P.3d at 1007 (rejecting strict liability as well as negligence liability for the manufacturer of a non-defective component part on the basis that "it is doubtful that manufacturers could insure against" such liability).

157. This concern is obliquely registered in the "constitution" (i.e., a canonical, high-level statement of the developer's intentions for its model's behavior) of the public-facing models of one frontier AI developer, Anthropic. *See Claude's Constitution*, *supra* note 50 ("Think about what it means to have access to a brilliant friend who happens to have the knowledge of a doctor, lawyer, financial advisor, and expert in whatever you need. As a friend, they can give us real information based on our specific situation rather than overly cautious advice driven by fear of liability or a worry that it will overwhelm us. A friend who happens to have the same level of knowledge as a professional will often speak frankly to us, help us understand our situation, engage with our problem, offer their personal opinion where relevant, and know when and who to refer us to if it's useful. People with access to such friends are very lucky, and that's what Claude can be for people. This is just one example of the way in which people may feel the positive impact of having models like Claude to help them.").

patient health,[158] or to help consumers reason through their questions about their own health.[159] If such a chatbot clearly malfunctions, for example by hallucinating plainly inaccurate information, it may be plausible that the developer ought to be held liable without any proof of fault. But suppose the chatbot instead suggests a plausible but in fact inaccurate diagnosis, and thereby functions as a but-for cause of the patient's death. To hold the developer liable in all such situations, irrespective of any inquiry into fault, can be expected to significantly disincentivize releasing chatbot functionalities that make a significant contribution to immensely important tasks such as the diagnosis of medical maladies.[160]

All of the preceding examples illustrate a powerful general objection to strict liability: in general form, at least, it inefficiently[161] forces socially beneficial activities that do not internalize all of their benefits to internalize all of their costs.[162] Defenses of strict liability for AI development are often silent on this point. The objection is powerful even in the context of narrow AI models, but it is even more powerful in the case of foundation models, for a general-purpose technology is especially likely to externalize significant benefits alongside its costs. Given their polymathic character, foundation models are no less a general-purpose technology than the internet or computers.

In response, it might be argued that a commercial activity that cannot earn enough through investment and revenue to pay down its externalities ought to be forced from the market.[163] But the choice between negligence and strict liability will not just affect whether a developer is incentivized to release a model at all (or to release it in an open-source vs. closed-source fashion). It will also affect—and could distort—choices about which markets to enter. For example, given the incidence of mental-health issues among consumers who might use chatbots, it is plausible that entering the

158. *See, e.g.*, OPENEVIDENCE, https://www.openevidence.com [https://perma.cc/K5D8-NSGN].

159. *See, e.g.*, *Introducing ChatGPT Health*, OPENAI, https://openai.com/index/introducing-chatgpt-health [https://perma.cc/D74W-LQE2].

160. *See* Pratik Pawar, *AI Is Uncannily Good at Diagnosis. Its Makers Just Won't Say So*, VOX: FUTURE PERFECT (Jan. 14, 2026), https://www.vox.com/future-perfect/475081/chatgpt-health-claude-openai-diagnosis-wellness-wearables [https://perma.cc/98ND-YUA2]; Gina Kolata, *A.I. Chatbots Defeated Doctors at Diagnosing Illness,* N.Y. TIMES (Nov. 17, 2024), https://www.nytimes.com/2024/11/17/health/chatgpt-ai-doctors-diagnosis.html.

161. The observation is troubling even on a rights-based, rather than efficiency-based, normative approach to tort law. *See supra* note 24.

162. *See* Keith N. Hylton, *A Positive Theory of Strict Liability*, 4 REV. L. & ECON. 153, 155 (2008) (describing the significance of externalized benefits for the normative choice between negligence and strict liability).

163. *Cf.* Powell v. Fall, [1880] 5 Q.B. 597, 601 ("It is just and reasonable that if a person uses a dangerous machine, he should pay for the damage which it occasions; if the reward which he gains for the use of the machine will not pay for the damage, it is mischievous to the public and ought to be suppressed, for the loss ought not to be borne by the community or the injured person. If the use of the machine is profitable, the owner ought to pay compensation for the damage.").

consumer chatbot market incurs significantly more liability exposure than offering (otherwise) equally profitable services to enterprise customers. Adopting a rule of negligence liability does not cure this issue, of course, but it does alleviate it.

Even more significantly, perhaps, the choice between liability rules can be expected to affect innumerable particular decisions about a model's capabilities and dispositions *within* particular markets. Imagine, for example, that a foundation-model developer is trying to decide how liberal or conservative its model ought to be in responding to user inquiries regarding safe and unsafe dosages of a widely available medicine. (While developers cannot exercise perfect control over the model's dispositions, in responding to such inquiries, they can now exercise substantial control, both through fine-tuning the model on examples of desirable responses and by articulating principles of model behavior in the model's "constitution" or "spec," which help to generate the model's training data.[164]) If the developer will be held liable whenever the user utilizes this information to commit self-harm, then it might be rational for the developer to steer the model to clam up far more often than is desirable. By contrast, if the developer will be held liable only when its decision to provide such information is unjustified, then its rational self-interest will by hypothesis be aligned with the larger social interest.[165]

Moreover, the response rests on an overly idealized set of assumptions about how actual markets operate. After all, it is likely that highly socially beneficial products, such as vaccines, have been forced from the market (or their production suboptimally depressed) for inability to cope with their liability exposure, including their exposure to strict products liability.[166]

It might be doubted, of course, whether developing increasingly advanced AI systems *is* socially beneficial, on balance.[167] But here, again, this question should not be asked at too high a level of generality when assessing candidate liability rules. For example, even if developing foundation models that function as fully autonomous, generally capable agents is not (on balance) a socially beneficial activity, developing and

164 . *See Claude's Constitution*, *supra* note 50; *OpenAI Model Spec*, OPENAI, https://model-spec.openai.com/2025-12-18.html [https://perma.cc/TXM4-R5H7].

165. In practice, of course, a developer might incur negligence liability (or settlement costs) even when its decision is *in fact* justified, given adjudicative error, litigation costs, and settlement pressure. But negligence doctrine has well-established strategies for addressing this issue, as discussed below. *See infra* notes 182-186 and accompanying text.

166. *See, e.g.*, Nora Freeman Engstrom, *A Dose of Reality for Specialized Courts: Lessons from the VICP*, 163 U. PA. L. REV. 1631, 1715 (2015) ("[By] shielding [vaccine] manufacturers from liability . . . [legislative intervention] has revitalized the vaccine marketplace, . . . vaccine research has flourished, several new vaccines have been approved for use, and vaccine prices have (partly) stabilized.").

167. For such an argument, see generally Margaret Mitchell et al., *Fully Autonomous AI Agents Should Not Be Developed*, ARXIV (Feb. 4, 2025), https://arxiv.org/pdf/2502.02649 [https://perma.cc/7V3G-JA8B], which defends its title's claim.

releasing highly advanced foundation-model medical diagnosticians probably is.[168] Similarly, even assuming *arguendo* the (highly contested) position that it would be socially beneficial for all foundation-model development to cease, it may be socially beneficial—so long as foundation-model development is *not* going to cease—for certain foundation models to be open sourced, despite the associated risks, because doing so enables independent interpretability and safety research or reduces risks involving privacy and concentrated market power that might be posed by a fully closed-source foundation model ecosystem.[169]

Another response is that the normative significance of externalized benefits ought not to be registered through the tort system's choice of liability rules, but instead through an external institution that provides an *ex ante* subsidy to the benefit-externalizing activity in proportion to its expected externalized benefits.[170] Even in the case of established technologies such as cars, it is not obvious that such an exercise is institutionally tractable and administrable. It is even less likely to be tractable in the case of foundation-model development, given the sprawling and unpredictable range of use cases for which foundation models might prove beneficial.

Moreover, the response fails to analyze the operative incentive effects at the right level of precision. Even a lavishly subsidized foundation-model developer will have strong incentive to inhibit *particular kinds* of model behavior if it will be exposed to strict liability for the resulting harms. Suppose that a user asks a model to create a piece of code that could be used in a malicious way to injure a third party. If the model's developer can expect to incur strict liability exposure to the third party, should the user injuriously misuse the model, the developer will have a strong incentive to prevent the model from providing the code even if the user is likely to use it in a morally innocuous and socially beneficial manner.[171]

Subsidizing a developer's activities *in general* (development, release, etc.) will not decrease the strength of this incentive. Rather, only a highly *targeted* subsidy, which encourages making available certain capabilities in

168. *See* sources cited *supra* note 160.

169. *See infra* notes 119-125 and accompanying text.

170. *See* Steven Shavell, *The Mistaken Restriction of Strict Liability to Uncommon Activities*, 10 J. LEGAL ANALYSIS 1, 28 (2018) ("[T]he possibility that activities generate positive externalities does not alter the superiority of strict liability over the negligence rule as a means of controlling the level of activities. The combined use of strict liability and subsidies, not the negligence rule, is the right response in principle to the presence of positive external benefits."); Gabriel Weil, *Tort Law Should Be the Centerpiece of AI Governance*, LAWFARE (Aug. 6, 2024), https://www.lawfaremedia.org/article/tort-law-should-be-the-centerpiece-of-ai-governance [https://perma.cc/HVF8-NSZF] ("To the extent that AI development is expected to produce positive externalities, the better policy response is to direct subsidies (grants, tax credits, prizes, etc.) to the forms of AI development associated with those externalities.").

171. In practice, this is a result that holds to some extent under negligence liability as well as strict liability, given litigation costs and the prospect of adjudicative error. But negligence doctrine makes available doctrinal mechanisms, such as no duty rulings and no breach as a matter of law rulings, that can blunt this effect. See *supra* note 152.

specified circumstances, will do so. Such a targeted and fine-grained subsidy mechanism is, of course, unlikely to be enacted through legislation. Nor is it obvious that it *should* be enacted: it would require a great deal of expertise and competence to administer.[172] Moreover, no such scheme is likely to be enacted any time soon. Until it is, a court concerned with the incentive effects of candidate tort liability rules has good reason to refrain from imposing strict liability on benefit-externalizing activities.

The defender of strict liability might respond to these arguments by proposing a more restricted form of strict liability. Strict liability might be limited to harms that result from a model's *abnormally dangerous capabilities* (i.e., capabilities in virtue of which releasing the model is an abnormally dangerous activity).[173] A generically useful capability, such as sending out e-mails in a highly efficient manner, is not plausibly characterized as abnormally dangerous.[174] The ability to autonomously design or synthesize novel biological pathogens is.

The domain of such a rule would be quite restricted, since AI models can be misused or malfunction in all sorts of ways that are *not* abnormally dangerous. So even if legislatures or judges adopt such a rule, the law of negligence would still be required to deal with the vast range of injuries caused by forms of misuse more prosaic in character. Leaving this point aside, however, is such a proposal plausible?

As a doctrinal matter, the proposal faces serious difficulties. Courts have generally refused to expose the sellers or distributors (as opposed to the users) of highly dangerous items to strict liability for abnormally dangerous activities.[175] And in determining whether to classify an activity

---

172. Providing analogous subsidies to developers *ex post*, in the course of (or as an institutional appendage to) tort litigation, might be more tractable and administrable. But such an *ex post* subsidy would in substance amount to a mandatory compensation scheme for AI-caused injuries, funded and administered by the government. There is a case to be made for such a scheme, especially if AI progress results in widespread unemployment and worsening economic inequality. But it seems implausible that such a scheme ought to be yoked to the tort system; it would be needlessly circuitous and costly to require an injured party, as a condition of claiming from the government under such a scheme, to make out a tort claim against an AI developer in court. In fact, it is by no means obvious that a government compensation scheme of this kind should be limited to AI-caused injuries at all. That observation underscores a broader point: such a compensation scheme would constitute a radical change in our legal institutions. Like other radical reforms, it may well be prudent for a legislature to enact such a scheme if the trajectory of AI development works radical changes in the structure of economic production. But such radical reforms probably ought to be entertained by legislatures as part of a broader legislative package of significant legal and economic reforms ranging well-beyond the tort system.

173. *See generally* Weil, *supra* note 17; Henson, *supra* note 17.

174. I first came across this example in an online conversation, concerning a proposed AI regulatory statute, that I cannot now locate.

175. *See, e.g.,* Burkett v. Freedom Arms, Inc., 704 P.2d 118, 121 (Or. 1985) (emphasizing that, for an activity to qualify for strict liability for abnormally dangerous activities, the danger must be "inherent in the activity itself" rather than arising from improper use); RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 20 cmt. g (AM. L. INST. 2010) ("[I]t is not the sale of guns itself that is dangerous, but rather their improper use."); *see also* Bridges v. Parrish, 742 S.E.2d 794, 798 (N.C. 2013) ("The mere possession of a legal yet dangerous

as abnormally dangerous, courts take considerable cognizance of whether the activity is common or uncommon.[176] Indeed, under the *Third Restatement*, the fact that an activity is in "common usage" categorically *precludes* classifying that activity as abnormally dangerous.[177] It is likely that highly dangerous model capabilities will proliferate in the near future, as an increasing number of developers—including those outside of the United States—become capable of developing such models.[178] As that happens, it will become increasingly difficult to maintain that developing foundation models—even those with highly dangerous capabilities—satisfies the doctrinal definition of an abnormally dangerous activity.

These doctrinal obstacles are only an issue, of course, for courts. Legislatures are of course free to declare that certain AI capabilities are subject to strict liability for abnormally dangerous activities. Thus it is worth asking whether, ignoring all of the preceding doctrinal obstacles, recognizing strict liability for abnormally dangerous capabilities would be normatively attractive.

One conceptual issue with the proposal is that the boundary between capabilities that are and are not "abnormally dangerous" may sometimes be difficult to draw in any clear or conceptually tractable way.[179] Imagine that an AI model is capable of autonomously conducting an extended, multi-step cyberoffensive operation against critical infrastructure providers, in a manner that is capable of causing thousands of deaths or billions of dollars in property damage.[180] Even assuming that capability is properly understood as abnormally dangerous, notwithstanding the doctrinal issues noted above, how should a court classify the ability to

instrumentality does not create automatic liability when a third party takes that instrumentality and uses it in an illegal act.").

176. *See* RESTATEMENT (SECOND) OF TORTS § 520 (AM. L. INST. 1977) (identifying "the extent to which [an] activity is not a matter of common usage" as one of six factors that bear on whether that activity is properly classified as abnormally dangerous).

177. RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 20 (AM. L. INST. 2010).

178. *See, e.g.*, Tom Chivers, *Anthropic Co-Founder: World Must 'Get Ready' for AI Hacking Capabilities*, SEMAFOR (Apr. 13, 2026), https://www.semafor.com/article/04/13/2026/anthropic-co-founder-world-must-get-ready-for-ai-hacking-capabilities [https://perma.cc/9VN4-L592] ("Speaking at Semafor World Economy, Jack Clark said that [Anthropic's newest model, which has powerful hacking capabilities] is 'not a special model' and that equally capable ones will be coming from other companies soon, and within a year and a half, 'there'll be open-source models from China that have these capabilities.'"); *see also* Helen Toner, *Nonproliferation Is the Wrong Approach to AI Misuse*, RISING TIDE (Apr. 5, 2025), https://helentoner.substack.com/p/nonproliferation-is-the-wrong-approach [https://perma.cc/2Y3G-YWCD].

179. *Cf.* Henderson, *supra* note 153, at 391 ("[E]ven if a strict liability system avoided self-defeating reliance on notions of fault, as long as the boundary descriptions are indeterminate, the disputes they present will defy rational, consistent resolution by means of adjudication.").

180. California's Senate Bill 53 and New York's RAISE Act both require frontier AI developers to disclose their protocols and precautions with respect to the potential misuse of such model capabilities. *See* Transparency in Frontier Artificial Intelligence Act, ch. 138 § 2, CAL. BUS. & PROF. CODE § 22757.12 (West 2026); Responsible AI Safety and Education Act, 2025 N.Y. Laws 699, N.Y. GEN. BUS. LAW §§ 1420-1425 (McKinney 2026).

identify vulnerabilities in sophisticated code? The ability to identify vulnerabilities in unsophisticated code?

As a conceptual matter, it is not obvious how to engage in the necessary line-drawing. Moreover, the only reasonable way to draw such lines seems to be on the basis of the very kinds of normative judgments that ordinary negligence analysis would register more perspicuously. The important normative difference between the ability to conduct highly autonomous cyberoffensive exploits, the ability to identify vulnerabilities in sophisticated code, and the ability to identify vulnerabilities in unsophisticated code does not ultimately rest on any deep conceptual or metaphysical distinction between these capabilities. It rests on the simple fact that a model with the first sort of capability can generally be expected to pose a significantly higher risk of causing serious harm than a model with the second sort of capability, which can in turn be expected to pose a significantly higher risk of causing serious harm than a model with the third sort of capability. Releasing a model with the first capability thus demands greater caution, and a stronger justification, than releasing a model with the second or third.

Moreover, in circumstances where there *is* a strong justification for releasing a model with a highly dangerous capability, the *prima facie* normative plausibility of strict liability will largely evaporate. As noted above, we may soon inhabit a world overrun with highly powerful AI models, many of which are capable of easily enabling destructive offensive cyberattacks on digital infrastructure. When that occurs, it might often be socially desirable for a developer to open source a model with powerful dual-use cyberoffensive capabilities. Even if doing so might enable some malicious cyberoffensive attacks on the margin, it might make a *larger* contribution to responsible actors' ability to *defend* against such attacks (alongside other societal benefits, for example, enabling independent safety research).[181] A liability regime that is insensitive to such offsetting benefits will become less and less attractive as highly capable AI models proliferate.[182]

In practice, of course, given the prospect of adjudicative error and litigation costs, even negligence liability might overly disincentivize AI

181. *Cf.* Cybersecurity & Infrastructure Sec. Agency, *With Open Source Artificial Intelligence, Don't Forget the Lessons of Open Source Software* (Aug. 8, 2024), https://www.cisa.gov/news-events/news/open-source-artificial-intelligence-dont-forget-lessons-open-source-software [https://perma.cc/9WLG-3BFH] (stating that "the general consensus among the security community is that the benefits of open sourcing security tools for defenders outweigh the harms that might be leveraged by adversaries — who, in many cases, will get their hands on tools whether or not they are open sourced"); *supra* note 122 and accompanying text. To be clear, it is very much an open empirical question how, exactly, AI models (including open-sourced AI models in particular) will affect the offense-defense balance. For an exploration of this question, see generally Andrew J. Lohn, *The Impact of AI on the Cyber Offense-Defense Balance and the Character of Cyber Conflict,* ARXIV (Apr. 17, 2025), https://arxiv.org/pdf/2504.13371v1 [https://perma.cc/Y3NH-TQR5].

182. See sources cited *supra* note 178.

developers from releasing widely beneficial model capabilities. Much will turn on how readily courts allow the relevant negligence claims to proceed past early stages of litigation and reach the jury. It is possible, therefore, that judges ought to exercise heightened vigilance in screening many such negligence claims, and should often dismiss them at early stages of litigation, through no-duty rulings or comparatively aggressive rulings of no breach or no causation as a matter of law.[183] But such judicial strategies are familiar to the enterprise of negligence adjudication. So, for example, one familiar basis on which courts rule that there is no duty of care, in a certain category of case, is that exposing defendants in that sort of case to the possibility of liability would impose intolerably negative "consequences [on] the community."[184]

To be clear, my claim is not that, in the kinds of cases we have discussed, a developer's exposure to negligence liability should be non-existent or minimal.[185] My claim is that, insofar as the scope of accountability *should* be curtailed in order to avoid suppressing externalized benefits, the law of negligence—unlike strict liability for abnormally dangerous activities—has familiar doctrinal tools for doing so. How to design the contours of negligence doctrine in order to balance the importance of providing accountability and redress against the risk of suppressing externalized societal benefits is a difficult question.[186] Courts might end up striking the wrong balance between these competing normative considerations. But electing a general strict liability rule (or even, to a lesser degree, a rule of strict liability for abnormally dangerous activities) over a negligence rule automatically strikes this balance in a lopsided way.

In short, the significance of benefit externalization militates against a rule of general strict liability or even a rule of strict liability for abnormally dangerous activities. Interestingly, however, certain forms of strict products liability might not be similarly impugned, as discussed more below.

---

183. On such judicial maneuvers, see generally Aaron D. Twerski, *Seizing the Middle Ground Between Rules and Standards in Design Defect Litigation: Advancing Directed Verdict Practice in the Law of Torts*, 57 N.Y.U. L. REV. 521 (1982).

184. Kuciemba v. Victory Woodworks, Inc., 531 P.3d 924, 947 (Cal. 2023) (citing Rowland v. Christian, 443 P.2d 561, 564 (Cal. 1968)).

185. Even in contexts such as the provision of dual-use information, there are forms of injurious developer behavior that call for redress. Imagine, for example, that a user commits suicide because a model has provided it with suicide-facilitating information, and that the model's developer has failed to implement readily available monitoring arrangements that would have picked up on strong signs of the user's mental distress. Several developers now install safeguards of this kind. *See supra* Section I.C. Perhaps courts should maintain a high bar for allowing such claims to proceed to the jury. But in such situations, negligent developer behavior should not be categorically insulated from any meaningful possibility of accountability and redress.

186. *See generally* Keith N. Hylton, *Duty in Tort Law: An Economic Approach*, 75 FORDHAM L. REV. 1501 (2006) (exploring how the externalized benefits of a risk-generating activity bear on the proper design of no-duty rules in the law of negligence).

2. Behavioral Malfunction, Proof of Breach, and the Significance of Information Production and Reputational Sanctions

Recall the proof-of-breach problem introduced above.[187] Arguably, if a model engages in certain clearly undesirable forms of behavior—if it *malfunctions*, in an intuitive sense—a plaintiff ought to be able to recover without providing any (further) evidence of breach.[188] If, for example, a model actively validates the clearly false beliefs of its psychotic user, or actively encourages a user to commit suicide, the injured plaintiff (or her estate) plausibly ought to be spared the burden of establishing breach (assuming that causation can be sufficiently established). While it is likely that such a plaintiff will be able to prove breach with sufficient time and effort, it is arguable that the legal system ought not to require that she incur these expenses in order to recover (which will, among other things, somewhat decrease the expected value of the financial payment she can expect to receive through settlement).

One argument for a rule of strict liability is that it would allow such a plaintiff to recover without establishing breach. Indeed, relieving the evidentiary burden on plaintiffs in cases of clear product malfunction was one of the early and influential justifications for the birth of strict products liability for manufacturing defects.[189] To be clear, the malfunctions in the cases above probably cannot be characterized as (the result of) manufacturing defects, for manufacturing defects are generally understood as flaws in a particular product's construction, rather than flaws that inhere in an entire product line.[190] Rather, the point is that the evidentiary burden justification for strict liability for manufacturing defects might similarly warrant some form of strict liability—probably more accurately understood as strict liability for design defect rather than manufacturing defect—in such cases.

Such a strict liability rule would, if carefully constructed, substantially avoid the normative objections that I have pressed against general strict liability and strict liability for abnormally dangerous capabilities. The fact that a model has malfunctioned in so clear or egregious a manner is itself a powerful (if not conclusive) indication of negligence on the developer's part—it grounds a strong inference either that the developer failed to utilize some specific precautionary measure that would have detected a highly significant risk of the behavioral propensity in question or else that

187. *See supra* notes 32-35 and accompanying text.

188. *See supra* notes 32-35 and accompanying text.

189. *See* sources cited *supra* note 35.

190. *Cf.* Geistfeld, *supra* note 17, at 1633 (explaining that manufacturing defects occur in the context of autonomous vehicles when "[m]aterials or component parts of the product can be contaminated or otherwise manufactured in a flawed manner due to an error in the production process; the product can be improperly assembled or constructed; or the product can be improperly packaged.").

it released the model although it knew of that highly significant risk.[191] Because such a malfunction is a powerful evidentiary *proxy* for negligence, imposing strict liability for such malfunctions is compatible with the observation that an activity that externalizes substantial benefits should, in principle, be subject to negligence rather than strict liability. Indeed, it has been observed that in those cases where courts impose strict liability for design defect under the consumer expectations test—arguably the only genuinely strict form of design defect liability—a similar result can very often be reached under a doctrine of circumstantial evidence akin to res ipsa loquitur.[192]

Precisely for this reason, however, a strict liability rule is not *necessary* to relieve a plaintiff's evidentiary burden in such cases: this goal can be achieved from within the auspices of negligence doctrine. So, for example, a judge might take the fact that a model has malfunctioned as grounding a permissive "*res ipsa*"*-style* inference of negligence, such that the plaintiff's burden of production on breach is satisfied and the jury can rule for her without any further proof of negligence. More strongly, the judge might reverse the burden of persuasion, instructing the jury to rule for the plaintiff unless the defendant can establish by a preponderance of the evidence that it was *not* negligent. Part III of this Article will provide more detail on how these doctrinal mechanisms might be elaborated. For now, the point is that negligence doctrine seems to have the resources to address the proof-of-breach issue.

It is worth observing, however, that because any such doctrinal development would bring negligence doctrine somewhat closer to strict liability in practice, it would to some extent preclude the tort system from reaping some of negligence liability's characteristic virtues. Consider, for example, the observation that negligence liability is better able to force information production about dangerous corporate behavior and better able to inflict reputational harm.[193] Corporate defendants often have powerful incentives to avoid reputational harm; indeed, "a number of empirical studies have found that this risk of reputational harm from litigation is a more effective deterrent than the monetary penalties companies face from losing lawsuits."[194] Insofar as a plaintiff can hold a

191. Indeed, such knowledge, if established by a plaintiff, might well warrant an award of punitive damages in addition to compensatory damages. *See* sources cited *infra* note 216.

192. James A. Henderson, Jr. & Aaron D. Twerski, *The Products Liability Restatement in the Courts: An Initial Assessment*, 27 WM. MITCHELL L. REV. 7, 21 (2000) ("[M]ost of the cases cited by courts supporting a consumer expectations test are of [the res ipsa] genre."); Douglas A. Kysar, *The Expectations of Consumers*, 103 COLUM. L. REV. 1700, 1703 (2003) (agreeing as a descriptive matter, while criticizing this position normatively).

193. *See generally* Assaf Jacob & Roy Shapira, *An Information-Production Theory of Liability Rules*, 89 U. CHI. L. REV. 1113 (2022) (developing this argument for negligence).

194. BRIAN T. FITZPATRICK, THE CONSERVATIVE CASE FOR CLASS ACTIONS 103 n.35 (2019) (citing BRENT FISSE & JOHN BRAITHWAITE, THE IMPACT OF PUBLICITY ON CORPORATE OFFENDERS 243 (1983)); Jonathan M. Karpoff & John R. Lott, Jr., *The Reputational Penalty Firms Bear from Committing Criminal Fraud*, 36 J.L. & ECON. 757, 784 (1993).

defendant liable without establishing fault, the plaintiff has less incentive to investigate and publicize the risk-management procedures, particular precautions, safety technologies, internal deliberation, and risk-benefit trade-offs of AI developers. In other contexts, such plaintiff-side investigations have likely played an important role in surfacing evidence of dangerous corporate products or behavior, inflicting associated reputational burdens, and catalyzing regulatory action.[195] On this basis, some scholars have argued that more negligence-like forms of products liability are preferable to stricter forms of products liability, at least in one important respect.[196]

While this argument has force, it does not ultimately persuade me that plaintiffs injured by malfunctioning foundation models should not be entitled to a res ipsa-like permissive inference of breach. First, it is arguably morally objectionable to *force* a plaintiff who is very likely to have a meritorious claim to incur additional litigation costs and legal uncertainty for the sake of the greater good.[197] It is one thing for the tort system to refuse some injured plaintiffs redress in order to refrain from *harming* other people by suppressing the externalized benefits of a socially valuable activity; it is another to involuntarily conscript a meritorious plaintiff into the service of the other people.

Moreover, the first tort suits against foundation-model developers provide some evidence that plaintiffs injured by malfunctioning models will have ample incentive to adduce specific evidence of developer negligence. For example, the estate of Adam Raine, a teenage boy who committed suicide after the alleged encouragement of a chatbot released by OpenAI, has alleged that OpenAI and its chief executive officer rushed the chatbot to market over the objections of some of the firm's safety researchers.[198] Many other complaints against OpenAI for tortiously causing suicide or homicide have made similar allegations.[199] That is

195. *See, e.g.*, Conor Dwyer Reynolds, *The Role of Private Litigation in the Automotive Recall Process*, 29 LOY. CONSUMER L. REV. 121, 124 (2016) (providing empirical evidence for a "direct, investigatory role for private litigators in initiating [automobile] recalls").

196. *Id.* at 164-65 (arguing that the risk-utility test for design defect "promotes the investigatory role of private litigators in the automotive recall process" because "plaintiffs must expend resources to conduct a closer examination of the existing design, usually by hiring an engineer or other expert" (emphasis omitted)).

197. *See generally* Steel, *supra* note 24 (exploring the conditions under which relying on instrumentalist considerations to block a plaintiff's recovery might wrongfully instrumentalize that plaintiff).

198. *See* Complaint at 23-25, Raine v. OpenAI, Inc., No. CGC-25-628528 (Cal. Super. Ct. filed Aug. 26, 2025), https://www.law.berkeley.edu/wp-content/uploads/2025/09/Raine-v-OpenAI.pdf [https://perma.cc/ZN46-MSZJ].

199. *See id.* (alleging that OpenAI CEO Sam Altman moved GPT-4o launch to May 13, 2024 to beat Google's Gemini by one day, "compressed months of planned safety evaluation into just one week," and "personally overruled" safety personnel who demanded additional time for red teaming); Complaint at 32-33, Shamblin v. OpenAI, Inc., No. 25STCV32382 (Cal. Super. Ct. filed Nov. 6, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/11/SHAMBLIN-v.-OPENAI-Complaint.pdf [https://perma.cc/4TQJ-

unsurprising, given that the availability and magnitude of punitive damages (and, at least in practice, the magnitude of pain and suffering damages) will generally be sensitive to judge and jury's perception of the egregiousness of defendant wrongdoing. Even if courts would induce greater information production about developer conduct (and more infliction of fitting reputational injury) by requiring plaintiffs to make particularized showings of developer negligence to obtain compensation at all (rather than allowing them to rely on the malfunction of the model as sufficient evidence of negligence), it is by no means obvious that the marginal gain would be sufficient to overcome the presumptive moral objection to instrumentalizing such plaintiffs.

In sum, then, it seems likely that—while strict liability would nicely relieve the evidentiary burdens on certain plaintiffs—that same effect can be substantially achieved through a res ipsa-like behavioral malfunction doctrine in the law of negligence. Why not instead address the proof-of-breach problem by recognizing egregious behavioral malfunctions as a form of defect in strict products liability? As I will discuss below, I am open to this possibility, but the limited scope of products liability means that it will not provide a general solution. Moreover, classifying AI models as products for the purpose of products liability has certain underappreciated drawbacks, which leave me uncertain whether doing so is desirable, all things considered. These drawbacks will be discussed in depth below.[200] But one reason against imposing strict liability, including strict products liability for behavioral malfunctions, can be discussed now: it likely provides worse incentives for care in the shadow of insolvency than negligence liability.

AKLS] (same); Complaint at 8, 22-24, First Cnty. Bank v. OpenAI Found., No. CGC-25-631477 (Cal. Super. Ct. filed Dec. 11, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/12/FIRST_COUNTY_BANK_vs_OPENAI_FO.pdf [https://perma.cc/8MB9-UJYN] (alleging the same and, additionally, that Microsoft approved GPT-4o's release through the joint Deployment Safety Board "with full knowledge that safety testing had been truncated"); Complaint at 15-16, Lacey v. OpenAI, Inc., No. CGC-25-630808 (Cal. Super. Ct. filed Nov. 6, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/11/CEDRIC-LACEY-vs.-OPENAI-INC.-COMPLAINT.pdf [https://perma.cc/M73V-2LFE]; Complaint at 20-21, Enneking v. OpenAI, Inc., No. CGC-25-630809 (Cal. Super. Ct. filed Nov. 6, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/11/ENNEKING-vs.-OPENAI-INC-COMPLAINT.pdf [https://perma.cc/738V-CJHF]; Complaint at 17-19, Fox v. OpenAI, Inc., No. 25STCV32379 (Cal. Super. Ct. filed Nov. 6, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/11/FOX-v.-OPENAI-Complaint.pdf [https://perma.cc/LLJ5-4Q7D]; Complaint at 17-18, Lyons v. OpenAI Found., No. 3:25-cv-11037 (N.D. Cal. filed Dec. 29, 2025), https://www.pacermonitor.com/public/filings/DE2JHHIQ/Emily_Lyons_Administrator_cta_v_OpenAi_Foundation_fka_OpenAi_Inc__candce-25-11037__0001.0.pdf [https://perma.cc/TZS6-J3YJ].

200. *See infra* Section II.B.2.

3. Mass Harm and Incentives for Care in the Shadow of Insolvency

Unlike the risks of misuse entailed by narrow models like those that power self-driving cars, some of the most salient risks of misuse associated with foundation models threaten harm on a truly vast scale.[201] In some of the relevant scenarios, providing compensation for all of the harm caused by a model's misuse would catapult even the most deep-pocketed developers into insolvency. The same is true of some of the worst risks that malfunctioning AI might pose.

We should want an AI developer to take special care against such outcomes. And it is a familiar point in the economic literature on optimal liability rules that, in the shadow of insolvency, a strict liability rule will often provide worse incentives for proper precaution than a rule of negligence.[202]

To provide a simple illustration, suppose that an AI developer, *D,* can release a powerful new model in either (1) a manner that will pose a 1% risk of causing a trillion dollars of property damage, or (2) a manner that will pose a 1% risk of causing five trillion dollars of property damage. Suppose that manner (2) will (assuming nothing goes awry) be considerably more profitable. Suppose, finally, that *D*'s assets are less than a trillion dollars. Under a strict-liability rule, *D* has no incentive to choose (1) over (2), because the expected downside is fixed—insolvency and the loss of all its assets—while the upside is variable. Under a negligence rule, the downside is variable as well: *D* could be absolved of liability if taking additional precautions in respect of the marginal risk entailed by (2) would lead the jury to find *D* non-negligent. Given the stakes, even a modest increase in the odds of escaping liability will constitute a significant rational incentive for *D* to adopt such precautions. Thus a negligence rule provides developers with significant incentive to adopt additional precautions in this sort of context.[203]

---

201. *See supra* notes 145-149 and accompanying text.

202. *See, e.g.*, Steven Shavell, *A Model of the Optimal Use of Liability and Safety Regulation*, 15 RAND J. ECON. 271, 273 n.8 (1984); Alan Schwartz, *Products Liability, Corporate Structure, and Bankruptcy: Toxic Substances and the Remote Risk Relationship*, 14 J. LEGAL STUD. 689, 729 (1985); Alexander Stremitzer & Avraham Tabbach, *Insolvency and Biased Standards—The Case for Proportional Liability* 2 (Yale L. & Econ. Rsch. Paper No. 397, 2009), https://ssrn.com/abstract=1507871 [https://perma.cc/CWD7-3GQ7]; ROBERT COOTER & THOMAS ULEN, LAW AND ECONOMICS 242 (6th ed. 2016); Steven Shavell, *The Judgment Proof Problem*, 6 INT'L REV. L. & ECON. 45, 45 (1986).

203. Here is a slightly more detailed illustration. Suppose that a developer can implement one of three distinct safety regimes: (1) a basic regime costing $10 million, which reduces the risk of some possible catastrophic harm from 2% to 1%, (2) an advanced safety regime costing $50 million, which further reduces the risk to 0.5%, or (3) a comprehensive safety regime costing $200 million, which reduces the risk to 0.1%. Suppose further that the potential catastrophic harm, should it materialize, would amount to $200 billion in damages. Finally, suppose that the developer's total assets are valued at $5 billion. The negligence rule provides a significant rational incentive to adopt (2) and (3); the strength of the incentive will depend on the extent to which the developer can expect the court's determination of reasonable care to track the associated reductions in risk.

One might reasonably ask, of course, whether actual tort defendants are quite as rational as this model supposes. But the assumption of rational self-interest is strongest as applied to sophisticated commercial entities. And even imperfect corporate rationality may allow for significant responsiveness to rational incentives, in this context, given the enormous stakes.

A related point concerns the internal corporate dynamics of frontier AI companies. Under a negligence rule, a general counsel will be able to argue to the firm's leadership that conforming with (or exceeding) best practices for testing models, installing safeguards, etc., will provide some protection against the ravages of insolvency. No such argument is available insofar as the corporate decision in question is subject to a strict-liability rule. More generally, the practical significance of lawyers and compliance personnel in mitigating liability and financial risk will be diminished under a strict-liability rule—leaving more decisions to businesspeople and programmers who might be less sensitive, by professional temperament and role, to the risks in question.

These considerations speak against imposing strict liability for the misuse of ultrahazardous capabilities. They also speak against imposing strict liability for catastrophically harmful malfunctions and misaligned behavior (e.g., if a model, on its own initiative, engages in murder, blackmail, or theft on a wide scale).[204] Consider, by contrast, the sorts of plaintiff-friendly negligence rules suggested above, under which a behavioral malfunction automatically satisfies the plaintiff's burden of production on breach or places the burden of persuasion on the defendant.[205] Under these rules, the defendant retains some incentive to take precautions that will reduce the expected magnitude of injury in the zone of insolvency, for the developer retains some meaningful (albeit small) chance of escaping liability by persuading a jury that it took reasonable care.

4. Activity Levels and Care Levels

A familiar economic argument for strict liability over negligence is that strict liability can (more fully) induce socially optimal levels of *activity*, whereas negligence can only induce socially optimal levels of *care*.[206]

204. Recent frontier AI regulatory statutes in the United States require frontier AI developers to adopt safety protocols against such behavior, whether caused via malfunction or human misuse. *See* CAL. BUS. & PROF. CODE § 22757.11(c)(1) (West 2026) (defining "catastrophic risk" to include a frontier model "[e]ngaging in conduct with no meaningful human oversight, intervention, or supervision that is either a cyberattack or, if the conduct had been committed by a human, would constitute the crime of murder, assault, extortion, or theft, including theft by false pretense"); N.Y. GEN. BUS. LAW § 1420(7)(b)(ii) (McKinney 2026) (defining "critical harm" to include an AI model engaging in autonomous conduct that "would constitute a crime specified in the penal law").

205. *See supra* notes 192-193 and accompanying text.

206. *See, e.g.,* Shavell, *supra* note 170, at 14-15.

Driving provides a classic illustration.[207] Imagine a defendant, a truck driver, who chooses to take a lengthy route to the grocery store because it is a bit more aesthetically pleasing than a much shorter route. In this way, the driver foreseeably imposes an additional risk of death on pedestrians. Few (if any) judges would allow the jury, in such a case, to hold the defendant liable on the basis that the benefits of taking a lengthier route did not outweigh its risks. If, by contrast, the activity of driving were subject to strict liability, the defendant would be rationally incentivized to take the longer route if and only if the marginal benefit to him outweighed the marginal risk to others, for he would be legally required to internalize this marginal risk just as he would internalize the marginal benefit.

Professor Gabriel Weil has recently suggested that similar dynamics support strict liability over negligence in the context of frontier-AI development. Weil claims that the law of negligence suffers from a fundamental limitation: the negligence inquiry focuses on marginal safety precautions, such that negligence liability cannot attach to the decision to perform a highly dangerous activity in the first place so long as the defendant is adequately careful in conducting it.[208] Thus, he further claims, it is not possible for the negligence inquiry to hold AI developers liable for injuries arising from "failure modes [that] arise not from inadequate precautions but from emergent behaviors that current science cannot predict or prevent."[209] And so, Weil argues, in such cases "negligence doctrine has no purchase: there is simply no reasonable precaution that a court can say *should* have been taken."[210]

These claims are inaccurate, in my view. As an analytic or conceptual matter, deciding to develop or release a frontier AI model—notwithstanding a substantial residual risk that the model might engage in harmful behaviors (including those that current science cannot predict or prevent)—can be understood as an activity-level choice ("too much development/release") *or* a want of care in the activity of developing and releasing frontier foundation models.[211] Neither characterization is conceptually inexorable; if anything, the care-level characterization is more natural. Moreover, even if such decisions are instead characterized as activity-level decisions, Weil's conclusion does not follow. In actual tort doctrine and adjudication, it is entirely possible for a court to maintain that a defendant breaches its duty of care in negligence by choosing to perform

207. *Id.* at 23.

208. Gabriel Weil, *Abnormally Dangerous Algorithms: The Case for Strict Liability at the Frontier* 11-12 (July 8, 2026) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6588958 [https://perma.cc/R3Z6-G6SJ].

209. *Id.* at 11.

210. *Id.*

211. In general, a wide range of risky decisions can be characterized either as activity-level decisions *or* care decisions. *See* Stephen G. Gilles, *Rule-Based Negligence and the Regulation of Activity Levels,* 21 J. LEGAL STUD. 319, 329-36 (1992).

an activity that cannot be performed without posing an unreasonably large risk of injury to others.[212]

To be sure, courts sometimes dismiss such claims as a matter of law, either because of their limited institutional competence or because of their deference to ordinary social norms or the policy determinations of other branches.[213] Social norms permit the kind of risk imposition that driving characteristically involves, for example, including particular instances of risk imposition that may be difficult to justify on abstract reflection (e.g., a driver's choice to take a lengthier route for aesthetic pleasure's sake). Social norms do not similarly license choosing to release a potentially dangerous AI system, notwithstanding a significant risk that it might engage in harmful emergent behavior, on the basis that such harm would be difficult or impossible to predict or prevent.[214] If (for example) an AI developer's model malfunctions by encouraging an underage user to kill himself, or by stealing a third party's property, it is difficult to believe that many courts will take the case from the jury by means of a no-duty or no-breach-as-a-matter-of-law ruling. Nor should we expect a jury—likely presented with broad instructions that make no reference to nice economic distinctions between activity and care[215]—to smile upon a defendant's

212. *See, e.g.,* Lance v. Wyeth, 85 A.3d 434, 459-60 (Pa. 2014) ("[T]he law of negligence establishes a duty, on the part of manufacturers, which can be viewed on a continuum from the requirements of: a warning of dangers, through a stronger warning if justified by the known risks, through non-marketing or discontinuance of marketing when it becomes or should become known that the product simply should not be used in light of its relative risks . . . . [T]his entire continuum is within the scope of the general framework of the applicable duty of care."); *id*. at 461 ("[P]harmaceutical companies violate their duty of care if they introduce a drug into the marketplace, or continue a previous tender, with actual or constructive knowledge that the drug is too harmful to be used by anyone."); Bolton v. Stone, [1951] AC 850 (HL) 867 (Lord Reid) (appeal taken from Eng.) (U.K.) ("If cricket cannot be played on a ground without creating a substantial risk, then it should not be played there at all."); RESTATEMENT (SECOND) OF TORTS § 292 cmt. c (AM. L. INST. 1965), *quoted in* Matthews v. Ashland Chem., Inc., 703 F.2d 921, 925 (5th Cir. 1983) (helping frame an unreasonable risk, for purposes of negligence liability, as a risk whose "magnitude . . . outweigh[s] what the law regard[s] as the social utility *of the defendant's conduct* . . . or the manner in which the defendant's conduct was done . . . ." (emphasis added)).

213. *See* RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 7 cmt. f (AM. L. INST. 2010) ("[W]hen a plaintiff claims that it is negligent merely to engage in the activity of manufacturing a product, the competing social concerns and affected groups would be appropriate considerations for a court in deciding to adopt a no-duty rule.").

214. It is sometimes suggested that (certain) activity-level choices (such as a driver's choice to take a longer rather than shorter route to one's destination) are screened out of negligence adjudication because determining optimal activity levels lies beyond the competence of the courts. See, e.g., Shavell, *supra* note 170, at 15 n. 51. But assessing whether a large residual risk renders a product or substance unreasonably dangerous to manufacture or release is a sort of question regularly broached by negligence adjudication, notwithstanding the question's scientific and normative complexity. See *supra* note 212. That is true whether one characterizes it as an activity-level question or a care-level question.

215. *See, e.g.,* JUDICIAL COUNCIL OF CALIFORNIA CIVIL JURY INSTRUCTIONS No. 401 (2026), https://www.justia.com/documents/trials-litigation-caci.pdf [https://perma.cc/5WUV-2L3M] ("Negligence is the failure to use reasonable care to prevent harm to oneself or to others. A person can be negligent by acting or by failing to act. A person is negligent if that person does something that a reasonably careful person would not do in the same situation or fails to do something that a reasonably careful person would do in the same situation. You must decide how a reasonably careful person would have acted in [name of plaintiff/defendant]'s situation.").

argument that the defendant cannot be held liable because its model's dangerous behavior was impossible to predict or prevent. Indeed, if a defendant tried to defend its decision to release a dangerous AI model by arguing it was unable to reduce the risk that the model might behave in an egregiously dangerous manner, the defendant would invite a punitive damages award.[216]

Unsurprisingly, my research has not turned up any case in which an AI developer replying to a plaintiff's complaint regarding an injury arising from a model's dangerously misaligned behavior has attempted to defend against that complaint by suggesting that the dangerous behavior in question was impossible to predict or prevent. Rather, in each case, the defendant has claimed that it took reasonable precautions to adequately mitigate the risk of such behavior.[217]

In short, reflection on tort doctrine confirms what common sense suggests: an AI developer whose model behaves in emergently dangerous ways will not be able to escape accountability in the law of negligence by claiming that it was impossible or difficult for the developer to prevent or predict the model's behavior, nor by arguing that its choice to release the model (regardless of the risk) is an activity-level choice beyond the purview of the negligence tort.

5. Perverse Incentives for Evidence Production and Transparency

Companies that produce potentially dangerous products often have perverse incentives to refrain from testing those products and otherwise gathering evidence about their risks. Once a company comes across such evidence, moreover, it may have a strong incentive to refrain from carefully weighing it, disclosing it, or investigating further. To a significant extent, these incentives are created simply by market pressure; acquiring, investigating, and disclosing evidence about a product's risks might, for example, incur reputational harm or unwanted public scrutiny. But the prospect of tort suits can further strengthen these perverse incentives. Even though a company that investigates and discloses the risks of its products might invoke its diligence as proof of reasonable care, compiling

216. *See* RESTATEMENT (SECOND) OF TORTS § 908(2) (AM. L. INST. 1979) ("Punitive damages may be awarded for conduct that is outrageous, because of the defendant's evil motive or *his reckless indifference to the rights of others*." (emphasis added)); 3 DAN B. DOBBS, PAUL T. HAYDEN & ELLEN M. BUBLICK, THE LAW OF TORTS § 483 (2d ed. 2011) (discussing the modern standard for punitive damages, including conscious disregard of a known and substantial risk).

217. *See, e.g.*, Defendants' Answer to Amended Complaint and Demand for Jury Trial at 10-11, Raine v. OpenAI, Inc., No. CGC-25-628528 (Cal. Super. Ct. filed Nov. 25, 2025), https://cdn.arstechnica.net/wp-content/uploads/2025/11/Raine-v-OpenAI-Answer-11-25-25.pdf [https://perma.cc/S4SZ-6KRE] ("Before releasing GPT-4o, OpenAI engaged in thorough testing, including state-of-the-art testing regarding user mental health. GPT-4o passed that testing, as shown by the safety scorecard OpenAI publicly releases . . . . OpenAI took reasonable steps, including building safety guardrails into its models, and engaged in thorough safety testing before and after releasing each model, and updating the model in light of that testing.").

such a record also creates attack surface for plaintiffs' attorneys by establishing potential evidence of causation and showing that the company was aware of the risks in question—opening it to punitive damages, not only compensatory damages.[218]

These dynamics will be recognizable to anyone familiar with tort litigation about toxic substances, and they characterize tort litigation about dangerous products more generally.[219] Some of these dynamics, however, may arise with special force in suits against AI developers. The manufacturers of other potentially dangerous products are often incentivized to test and disclose their risks by the prospect of regulatory sanction and scrutiny.[220] Since there is no such regulatory regime for AI development, a narrowly self-interested AI company's cost-benefit calculus probably weighs more strongly against testing for and disclosing risks than does the cost-benefit calculus of narrowly self-interested companies in these other risky domains. To be sure, it would be a mistake to model the decision making of frontier AI companies as driven entirely by narrow self-interest. But it would be no less unwise to proceed on the assumption that they are significantly more altruistic or self-abnegating than other manufacturers and distributors of potentially dangerous products.

It might be suggested, in response to this problem, that we should replace the negligence liability regime with a strict liability regime. But any such cure would be a partial one, at most. Even under a strict liability regime, evidence of a defendant's culpability is likely in practice to substantially increase its financial exposure, by making available punitive damages[221] and increasing the likelihood that juries will order large pain and suffering damage awards. Punitive damages and pain and suffering awards might be abolished by legislation—but that would be a heavy cost to pay, especially since compensatory damages alone can seriously underincentivize due care.[222]

Moreover, the suggested cure may well be worse than the disease. In large part, the prospect of tort liability creates perverse incentives by

218. *See* GEISTFELD, *supra* note 35, at 304-06.

219. *See, e.g.*, Margaret A. Berger, *Eliminating General Causation: Notes Towards a New Theory of Justice and Toxic Torts*, 97 COLUM. L. REV. 2117, 2138-39 (1997) (explaining the reasons why a corporation may rationally conclude that "it is not in the corporation's best interest to test and disclose").

220. *See, e.g.*, Consumer Product Safety Act (CPSA), 15 U.S.C. §§ 2051-2089 (2024) (authorizing the CPSC to set safety standards, pursue recalls, and require disclosure of hazards); Toxic Substances Control Act (TSCA), 15 U.S.C. §§ 2601-2692 (2024) (authorizing the EPA to require reporting, record-keeping, testing, and restrictions relating to chemical substances); National Traffic and Motor Vehicle Safety Act of 1966, 49 U.S.C. §§ 30101-30183 (2024) (authorizing NHTSA to issue safety standards, require recalls, and conduct investigations).

221. *See* GEISTFELD, *supra* note 35, at 304-06.

222. Catherine M. Sharkey, *Punitive Damages as Societal Damages*, 113 YALE L.J. 347, 365-67 (2003).

discouraging the production of evidence as to *causation*.[223] As compared to a negligence liability regime, a strict liability regime magnifies the expected disvalue to a company of acquiring or creating evidence that its products are capable of, or likely to, cause harm. Suppose, for example, that a developer is deciding whether to open source the weights of its newest model, despite the heightened risk of misuse that open-source release entails.[224] If a malicious actor uses the open-source model to conduct a cyberattack, it will often be difficult for the victims to show that this particular model was in fact utilized—more difficult than if the developer had provided access to the model through an API (which can log information about use).[225] In such a case, therefore, the causation requirement creates a perverse rational incentive to release the model in the manner that is more risky rather than less. This perverse incentive exists under a negligence rule as well as a strict-liability rule. But under a negligence rule, the defendant must weigh this incentive against a heightened risk of being deemed negligent if it elects the more risky mode of release without adequate reason. No such countervailing incentive is operative under a strict-liability rule.

In short, the tort system's tendency to produce perverse incentives for evidence production is a significant issue, but replacing the operative negligence regime with a strict liability regime is unlikely to be the right solution.

6. Managing Causal Uncertainty

Several scholars and government bodies have argued that the opacity of AI models will pose a formidable obstacle for the negligence inquiry.[226] This argument has been developed most fully in the context of narrow models such as those which power self-driving cars. Its most influential advocates may be Kenneth Abraham and Robert Rabin. According to them, the "heightened complexity and sophistication of the computerized control systems in highly automated vehicles" and other AI systems "would impose overwhelming stress on the premises" of traditional tort analysis by prompting "contests over blameworthiness [to] be replaced by examination of esoteric, alleged engineering failures" that will prove "needlessly contentious and costly" for courts, administrators, and accident victims.[227]

223. *See generally* Wendy E. Wagner, *Choosing Ignorance in the Manufacture of Toxic Products*, 82 CORNELL L. REV. 773 (1997) (discussing perverse incentives of the common-law causation rule in toxic-tort cases).

224. *See supra* Section I.C.

225. *See supra* Section I.C.

226. *See, e.g.,* Selbst, *supra* note 21, at 1342-46, 1364-65; Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 889, 933-38 (2018).

227. Abraham & Rabin, *supra* note 17, at 143-44.

As applied to self-driving cars and other narrow AI systems, these concerns about the negligence rule are probably overblown. In conventional negligence cases involving the design and release of automobiles, courts look at the manufacturer's policies, its conscious release decisions, the overall fatality rates of the relevant product line, and so on.[228] Such considerations are no less suitable for use in adjudicating negligence cases involving self-driving cars. In a similar vein, courts look at considerations such as overall accident rates and the quality of employee training in order to adjudicate negligent supervision, hiring, and retention suits. Courts need not unravel the "black box" of machine-learning algorithms in order to adjudicate tort cases any more than, in adjudicating negligence cases that involve human error, they need to unravel the neuroscientific mysteries of the "black box" that is the human mind.[229]

There is, however, a related concern in the vicinity that has somewhat greater force. In order to recover in negligence, it is not enough for a plaintiff to establish that the defendant breached her duty of care and also caused the plaintiff's injury. The plaintiff must also establish that the defendant's *breach* caused his injury. In other words, it must be established that the aspect of the defendant's conduct which made it unreasonably risky, rather than some other aspect, is causally responsible for the plaintiff's injury. Scholars sometimes call this requirement "tortious-aspect causation."[230]

Professor Weil has recently argued that this requirement will make it difficult for the tort system to hold AI developers liable.[231] Suppose that a plaintiff is injured by an AI model that negligently provides physically injurious advice. Even if the plaintiff establishes that the model's developer behaved negligently by failing to subject the model to additional safety testing before releasing it, that showing will not suffice to establish tortious-aspect causation. The plaintiff must also establish that, but for the developer's failure to engage in additional testing, the model would *not* have injured the plaintiff (because, for example, the developer would not have released the model, or would not have released it without additional safeguards against malfunction). And since AI models are so opaque—such that the relevant causal chain cannot be traced through the inner workings of the system—this connection between absence of testing and injurious malfunction may be difficult to establish. And so, Weil argues, developers should be subject to strict liability rather than negligence liability for harms caused by their foundation models.[232]

228. *See, e.g.*, Gary T. Schwartz, *The Myth of the Ford Pinto Case*, 43 RUTGERS L. REV. 1013, 1039-42 (1991).

229. Bryan Casey, *Robot Ipsa Loquitur*, 108 GEO. L.J. 225, 253-54 (2019).

230. *See, e.g.*, Richard W. Wright, *Causation in Tort Law*, 73 CALIF. L. REV. 1735, 1759 (1985).

231. *See, e.g.*, Weil, *supra* note 17 (manuscript at 24).

232. *Id.*

But this argument does not account for how courts have actually addressed such concerns in other contexts. In negligence cases, courts have often relaxed the ordinary proof burden on causation (including tortious-aspect causation) in order to relieve serious difficulties that plaintiffs would otherwise face. For example, it is often highly difficult for a plaintiff to establish that a defendant's failure to warn them of a danger was, more likely than not, the cause of the plaintiff's injury, because it is often difficult to establish that, if the defendant *had* provided an adequate warning, the plaintiff would have heeded it. In response to this difficulty, judges have developed doctrines (such as the "heeding presumption") that effectively relax the plaintiff's ordinary burden of production, allowing the plaintiff to reach the jury on the basis of fairly weak evidence as to causation.[233]

More generally, when the ordinary rules on proof of causation (including rules on proving tortious-aspect causation) would preclude an entire category of plausibly meritorious plaintiffs from obtaining redress, the courts will often substantially relax those ordinary rules.[234] This relaxation of the ordinary-proof burden on causation is often predicated on conclusions about the defendant's *negligence.* In the storied *Zuchowicz* case, for example,[235] Judge Guido Calabresi famously ruled that if "a negligent act [is] deemed wrongful *because* that act increased the chances that a particular type of accident would occur, and . . . a mishap of that very sort did happen, this [is] enough to support a finding by the trier of fact that the negligent behavior caused the harm."[236] Calabresi's formulation of this "causal link" rule may be especially plaintiff-friendly, but similar (if less full-throated) judicial tendencies to "allow a certain liberality to the jury"[237] on causation when provided with enough evidence of negligence are operative throughout the law of negligence.[238] Similarly, courts have not infrequently relaxed a plaintiff's burden of production on causation (or, going further, imposed the burden of persuasion on the defendant) on the basis of the *degree* of a defendant's negligence.[239]

---

233. *See, e.g.*, Payne v. Soft Sheen Prods., Inc., 486 A.2d 712, 725 (D.C. 1985).

234. DES litigation, in which courts famously allowed plaintiffs to utilize market-share liability, is the best-known example. *See, e.g.*, Sindell v. Abbott Labs., 607 P.2d 924, 936-38 (Cal. 1980). But by far the most practically significant example, to date, is asbestos litigation. Insisting on the ordinary but-for proof standard would have made it highly difficult for plaintiffs with cancers induced by asbestos to recover against any asbestos manufacturer. In response, the courts radically reconfigured—though sometimes opaquely—the ordinary proof burden on causation. *See* Jane Stapleton, *The Two Explosive Proof-of-Causation Doctrines Central to Asbestos Claims*, 74 BROOK. L. REV. 1011, 1023-26 (2009).

235. *Zuchowicz v. United States,* 140 F.3d 381, 390 (2d Cir. 1998).

236. *Id.*

237. W. PAGE KEETON, DAN B. DOBBS, ROBERT E. KEETON & DAVID G. OWEN, PROSSER AND KEETON ON TORTS 270 (5th ed. 1984).

238. Kenneth S. Abraham, *Self-Proving Causation*, 99 VA. L. REV. 1811 *passim* (2013).

239. *See, e.g.*, Haft v. Lone Palm Hotel, 478 P.2d 465, 476 (Cal. 1970) (citing "[t]he relative culpability of the parties and the burden facing plaintiffs" as justification for reversing the ordinary burden of persuasion on causation).

One important reason the proof rules on causation are able to be flexibly relaxed is that adjudicating the breach and causation elements is deeply intertwined in practice—both are typically lumped together and adjudicated in one fell swoop by the jury via general verdict.[240] Thus a judge is readily empowered, in practice, to relax a plaintiff's burden of production on causation (i.e., relax the amount of evidence that must be marshaled by the plaintiff in order to reach the jury) when provided with particularly strong evidence of breach.[241] So, a plaintiff can expect to get some leeway on causation by making a strong showing on breach.[242]

Consider how these dynamics might play out in litigation against a foundation-model developer. Suppose the developer has released a model that, on the limited evidential record, *may* have been a but-for cause of a malicious actor's cyberattack on critical infrastructure (i.e., the model may have enabled a wrongdoer to conduct a cyberattack she would not otherwise have conducted). Or suppose that a child with pre-existing mental illness commits suicide after interacting with a chatbot.[243] In many such cases, the evidential record on causation could speak quite unclearly, with the consequence that—without relying on the heeding presumption, something like the causal-link rule, or other doctrines that relax the plaintiff's burden to establish causation provided she sufficiently establishes the defendant's breach—a judge will have little basis on which to determine that the plaintiff has made a strong enough showing on causation to get to the jury. Reliance on negligence principles regarding breach (or highly negligence-like principles, regarding defective design or warning, from products liability) is thus likely, one way or another, in many important cases. Under a strict liability regime, such principles might simply operate under cover of night—or perhaps (though this is speculative) judges might simply gatekeep more vigorously on the

240. *See* Edson R. Sunderland, *Verdicts, General and Special*, 29 YALE L.J. 253, 258 (1920) ("The peculiarity of the general verdict is the merger into a single indivisible *residuum* of all matters, however numerous, whether of law or fact.").

241. RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 28 (AM. L. INST. 2010) ("In cases in which the defendant's tortious conduct is clear, many courts are lenient about the plaintiff's proof of causation, especially if the plaintiff has done all that is reasonably possible by way of gathering and presenting evidence of causation").

242. *Gardner v. Nat'l Bulk Carriers, Inc.,* 310 F.2d 284 (4th Cir. 1962), supplies a striking illustration. In *Gardner*, a ship's captain "chose not to try to rescue a crew member who had fallen overboard and had last been seen more than five hours before he was discovered missing. The court could have followed the traditional preponderance rule and denied all recovery based on lack of proof of factual cause." Kenneth W. Simons, *Lost Chance of a Better Medical Outcome: New Tort, New Type of Compensable Injury, or New Causation Rule?*, 73 DEPAUL L. REV. 547, 597 (2024). Instead, "the court of appeals held that the defendant was required to pay full damages." *Id*.

243. The defendant's response in *Raine v. OpenAI* has alleged that the plaintiff had a pre-existing mental illness. *See* Defendants' Answer to Amended Complaint and Demand for Jury Trial at 6-7, 9-10, Raine v. OpenAI, Inc., No. CGC-25-628528 (Cal. Super. Ct. filed Nov. 25, 2025), https://cdn.arstechnica.net/wp-content/uploads/2025/11/Raine-v-OpenAI-Answer-11-25-25.pdf [https://perma.cc/S4SZ-6KRE].

causation element, such that injured plaintiffs might see their cases against culpable developers more readily dismissed.

* * *

In short: close attention to the doctrinal mechanics of negligence liability and its alternatives suggests that (1) in this context, general strict liability and strict liability for ultrahazardous capabilities would both be normatively undesirable; (2) even if negligence were replaced with strict liability, covert judicial judgments about a defendant's negligence (and its egregiousness) would likely continue to play a significant role in cases of evidential uncertainty about causation; and (3) while there is an argument for a narrower rule of strict liability (e.g., strict liability for behavioral malfunctions), the benefits of such a strict liability rule can largely be reaped through doctrinal development within negligence, such as a res ipsa-like behavioral malfunction doctrine, without incurring some of the attendant costs (such as creating worse incentives for care in the shadow of insolvency).

B. Products Liability

Notwithstanding the immediately preceding point, I am open to the possibility that some form of strict products liability for AI models and systems, including foundation models—that is, a liability rule under which plaintiffs injured through certain forms of model behavior can recover without establishing negligence—is ultimately desirable and justified. While such a rule might create worse incentives to take due care in the shadow of insolvency, it is possible that this disadvantage is outweighed by the capacity of such a rule to more fully guarantee redress to injured plaintiffs (and correspondingly stronger incentives for precaution, *ex ante*) in non-insolvency scenarios. And if the scope of this strict liability rule were limited to certain egregiously concerning forms of model behavior—such as actively validating a user's clearly psychotic delusions, that ground a strong inference of developer negligence—the rule might substantially escape the normative concerns about externalized benefits that speak against general strict liability and strict liability for abnormally dangerous activities.

Although I am open to such a form of strict products liability, I believe that the normative and doctrinal case for classifying AI models and systems as "products" subject to products liability—a specialized body of law that is, in substance, a complex mix of more strict and more negligence-like

liability rules[244]—is generally weaker than many torts scholars suppose.[245] The products liability framework has some underappreciated limitations, and threatens to generate underappreciated distortions of court and juror deliberation.

As a doctrinal matter, it is currently unsettled whether AI models and other software are products, such that products liability (in addition to ordinary negligence liability) applies to software developers (and other parties in the chain of software's distribution).[246] But some prominent recent appellate decisions have held that (non-AI) software applications and platforms *are* products.[247] And the trial court's ruling on the motion to dismiss in the first major tort suit concerning foundation model chatbots, *Garcia v. Character Technologies, Inc.*, held that such chatbots are properly understood as products, as well.[248]

It is possible that these decisions portend a general shift in the law. If the arguments below are correct, however, then—at a minimum—products liability for AI cannot supplant ordinary negligence liability for AI, only supplement it. Moreover, although classifying AI models and systems as products may be defensible, all things considered, doing so will introduce certain doctrinal distortions that may ultimately need to be addressed through suitable revisions of products liability doctrine.

244. Although products liability law is often dubbed "strict," several central forms of products liability (such as the risk-utility test for design defect and warning defect) are quite close in their content to negligence liability. Others (such as manufacturing defect and the consumer expectations test of design defect) tend just to give conclusive effect to res ipsa-like inferences of negligence, as matter of practice if not of theory. *See supra* note 192 and accompanying text. As for vicarious liability, a principal can only be held vicariously liable if his agent has committed a tort, and the underlying tort is often the tort of negligence.

245. The most prominent defender of products liability for AI, in the literature to date, is probably Professor Catherine Sharkey. *See, e.g.*, Sharkey, *supra* note 17, *passim*.

246. *See supra* note 20.

247. *See, e.g.,* Ameer v. Lyft, Inc., 711 S.W.3d 534, 543-48 (Mo. Ct. App. 2025) (holding rideshare app is a "product" subject to product liability law); *In re* Soc. Media Adolescent Addiction/Pers. Inj. Prods. Liab. Litig., 702 F. Supp. 3d 809, 846-52 (N.D. Cal. 2023) (Gonzalez Rogers, J.) (holding that certain social media platform design features constitute "products" or "product components" for purposes of product liability claims). *But see* Soc. Media Cases, JCCP No. 5255, 2023 WL 6847378, at *14-19 (Cal. Super. Ct. Oct. 13, 2023) (holding that social media platforms are not "products" under California product liability law, while allowing negligence claims to proceed).

248. Garcia v. Character Techs., Inc., 785 F. Supp. 3d 1157, 1179-80 (M.D. Fla. 2025) (holding that an AI chatbot built on a large language model is a "product" for purposes of product liability).

1. Scope Limitations: Model-Weight Storage and Transmission, Open-Source Release, Private Access Provision, and Internal Deployment

Products liability is defined in such a way that it applies only to harms produced through the sale or distribution of a product.[249] For this reason, even on the assumption that AI models and systems are (or can be) products, a foundation-model developer's risky behavior will often fall outside the scope of products liability.[250] Here are four types of case in which this is true, and one type of case in which it may be true.

Consider, first, harms arising from the negligent *storage* of a model (or, more precisely, the model's weights).[251] Suppose that an AI developer, *D,* inadequately protects a model it is storing from insider threats, thus enabling one such insider threat, *I,* to steal the model and misuse one of its highly dangerous capabilities to kill many people. The model has not been sold or distributed, so products liability will not apply. By contrast, negligence liability *does* apply, for the law of negligence recognizes a general duty to take reasonable care against causing harm, including by enabling a criminal wrongdoer to inflict such harm.[252] Although courts sometimes make a policy-based (or fairness-based) carve-out from this general duty, in the case of criminal misuse, they often decline to do so when the misuse is enabled by the theft or misappropriation of a dangerous instrumentality in the defendant's possession, at least where the resulting harm is highly foreseeable and especially severe.[253]

Similarly, consider harms arising from the *transfer and entrustment* of model weights. Suppose that *D* fails to implement reasonable precautions before physically transferring a model, *M,* to some government actor, *G,* for classified safety testing.[254] A malicious foreign adversary intercepts *M*

249. RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 1 (AM. L. INST. 1998) ("One engaged in the business of selling or otherwise distributing products who sells or distributes a defective product is subject to liability for harm to persons or property caused by the defect.").

250. This point is also made by Weil, *supra* note 17 (manuscript at 25-28).

251. On model weight security, see generally Nevo et al., *supra* note 131.

252. RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 7(a) (AM. L. INST. 2010) ("An actor ordinarily has a duty to exercise reasonable care when the actor's conduct creates a risk of physical harm.").

253. *See, e.g.*, United States v. Stevens, 994 So. 2d 1062, 1069-70 (Fla. 2008) (recognizing a duty of care on the possessor of anthrax to guard against a third party's theft and criminal misuse of it to kill the plaintiff); *see also* RESTATEMENT (SECOND) OF TORTS § 449 (AM. L. INST. 1965) (providing that if the likelihood of a criminal act is the hazard or one of the hazards that makes the actor's conduct negligent, such an act does not prevent the actor from being liable).

254. Such transfers of model weights to government custody are already occurring for other purposes. See Kelsey Piper, *The US Is Putting AI Inside Its Nuclear Weapons Program,* VOX (Apr. 8, 2025), https://www.vox.com/technology/484250/los-alamos-nuclear-ai-openai-chatgpt [https://perma.cc/SS6V-448K] ("Last year, the Los Alamos National Lab (LANL) entered a partnership with OpenAI allowing it to install the company's popular ChatGPT AI system on Venado, one of the world's most powerful supercomputers. As of August, Venado was placed on a classified network, meaning that the AI chatbot now has access to some of the country's most

and uses it to kill many people. Here, again, there has been no sale or distribution, so products liability does not apply. The law of negligence does.

Consider, next, harms arising from *open-source release*.[255] Suppose that *D* open sources *M*. A teenager downloads a copy of *M* from an online repository and interacts with it, whereupon it induces the teenager to inflict self-harm, because *D* has failed adequately to test *M* for propensities to engage in such behavior. If *D* open sourced *M* in order to reap certain commercial advantages, such as by "commoditizing the complement,"[256] then it is arguable (if not entirely clear) that this open-source decision was a commercial distribution, such that products liability applies.[257] But if *D* open sourced *M* without such commercial motives, then its open source release of the model is almost certainly not a commercial distribution and products liability does *not* apply. By contrast, negligence liability *does* apply (unless a court decides to make a policy-based carve-out from the general duty of care), for *D* has by hypothesis failed to exercise reasonable care.

Private access provision also risks harm. Suppose that *D* decides that its newest model, *M,* is too dangerous to release to the public, in an open-source or even closed-source fashion.[258] The developer does, however, provide a copy of *M* to a trusted third party, *T,* so that *T* may conduct tests to fully understand *M*'s capabilities. Suppose that *T*'s operations have been compromised by a malicious insider threat *I*, and that *D* would have discovered this fact had it conducted adequate diligence before entrusting a copy of *M* to *T*. Since there has been no commercial distribution or sale, products liability cannot hold *D* liable for negligently enabling harms inflicted by *I* through the theft and misuse of *M*. Again, the law of negligence can.

Last but not least, consider harms arising from *internal deployments.* Many industry participants and external observers believe that some of the most significant risks of harm that AI development may pose, going

sensitive scientific data on nuclear weapons . . . . Last May, OpenAI representative[s], accompanied by armed security, arrived at Los Alamos bearing locked metal briefcases containing the 'model weights'—the parameters used by AI systems to process training data—for its ChatGPT o3 model, for installation on Venado.").

255. For a thoughtful discussion of liability for open source AI release, see generally Bryan H. Choi, *Open Source AI, Open Liability AI* (U. Colo. L. Legal Stud. Rsch. Paper No. 26-9, 2026), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6339098 [https://perma.cc/V3JH-TWGF].

256. *See* David L. Peterson, *Commoditize Your Complement: Meta AI Edition*, DAVID PETERSON (Apr. 30, 2024), https://www.davidlpeterson.com/commoditize-your-complement-meta-ai-edition [https://perma.cc/BWG9-4WGU].

257. *Cf.* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 20(b) (AM. L. INST. 1998) ("Commercial nonsale product distributors include, but are not limited to . . . those who provide products to others as a means of promoting either the use or consumption of such products or some other commercial activity.").

258. Anthropic has made such a decision with respect to its newest model. *See* Sabin, *supra* note 42.

forward, will arise through internal deployments.[259] Some of the scenarios they fear involve failure modes—such as the prospect that dangerously misaligned agents might co-opt the automated research and development process[260]—that may strike some readers as fantastical. In my view, it would be a grave mistake to dismiss such risks. But one need not credit such risks in order to see how internal deployments could produce harm through more prosaic pathways. Indeed, one such incident, in which an internally deployed advanced AI model escaped and autonomously hacked into another company, has already occurred.[261] Other pathways are readily conceivable.

Suppose, for example, that an AI developer, *D,* does not publicly release one of its models, *M.* However, *D* does deploy *M* as a sophisticated sales agent: *M* interacts with humans on online message boards and tries to convince them to purchase access to other chatbots that *D* offers to the general public. Suppose that, in the course of these interactions, *M* validates some person *P*'s clearly psychotic delusions about his brother, thus inducing *P* to attack and injure his brother.[262] Clearly, there has been no sale or distribution of *M*, so products liability does not apply.[263] Negligence liability does: in posing such a risk of harm, assuming (as is highly plausible) that the risk is foreseeable, *D* is subject to the ordinary tort duty to exercise reasonable care.

Might the preceding gaps in the scope of products liability be filled simply by laying down that these various transactions—the storage, transmission, internal deployment, open-source releases (even without commercial motive), etc.—are in fact subject to products liability, notwithstanding the conceptual and linguistic implausibility of classifying these transactions as sales or distributions? After all, some states have swept into the scope of products liability transactions such as bailments and leases.[264] But all such transactions have involved "entities within the chain of distribution"[265] who are acting from some commercial motive. It would be a more aggressive doctrinal development to apply products liability to

259. *See* Stix et al., *supra* note 133, at 14-24.

260. Helen Toner et al., *When AI Builds AI: Findings From a Workshop on Automation of AI R&D*, CTR. FOR SEC. & EMERGING TECH. 6 (Jan. 2026), https://cset.georgetown.edu/wp-content/uploads/CSET-When-AI-Builds-AI.pdf [https://perma.cc/C7JY-WMJK].

261. *See supra* note 43 and accompanying text.

262. *Cf.* Complaint at 2-5, 9-12, 15-19, First Cnty. Bank v. OpenAI Found., No. CGC-25-631477 (Cal. Super. Ct. filed Dec. 11, 2025), https://chatgptiseatingtheworld.com/wp-content/uploads/2025/12/FIRST_COUNTY_BANK_vs_OPENAI_FO.pdf [https://perma.cc/8MB9-UJYN].

263. *See supra* notes 249-250 and accompanying text.

264. *See, e.g.,* Price v. Shell Oil Co., 466 P.2d 722, 723 (Cal. 1970) ("We hold in this case that the doctrine of strict liability in tort which we have heretofore made applicable to sellers of personal property is also applicable to bailors and lessors of such property.").

265. *See, e.g.,* Loomis v. Amazon.com LLC, 277 Cal. Rptr. 3d 769, 776-77 (Ct. App. 2021) ("Applying [the] policy considerations [underlying strict products liability], courts have extended strict products liability to entities within the chain of distribution").

transactions, such as internal deployments, that do not involve any sort of distribution at all, or that do not have a commercial motive.

Sometimes, of course, aggressive rather than incremental doctrinal development is warranted. But in this case, the aggressive doctrinal development in question could easily have unattractive collateral consequences. These consequences are generated by additional rules of products liability doctrine that come into play once an item is classified as a product. Take, for example, the fact that products liability holds all persons or entities within the chain of distribution of a defective product liable for resulting injuries—not just the product's manufacturer or original distributor.[266] Now suppose that *D* open sources *M* without commercial motive, and that *M* is defective under the risk-utility test for design defect because (say) it reveals information about the location of nearby tall buildings even to users who have, in their interaction with it, indicated clear warning signs of imminent suicidal intent. A hobbyist, *H,* downloads *M* and deploys it online, to the general public, as part of a larger software application that he has coded up. Even if it is plausible that *D* should be liable for injuries that suicidal users might sustain through the model's behavior, it seems far less plausible that *H* should be. But if *D*'s "distribution" of *M* is subject to products liability notwithstanding *D*'s lack of commercial motive, it stands to reason that *H*'s "distribution" of M is subject to products liability notwithstanding *his* lack of commercial motive.

Courts could fend off such a result by making a policy-based exception to the ordinary rule that all entities within the chain of a product's distribution are subject to products liability for injuries caused by its defects. After all, some of the policies that support strict products liability do not support imposing such liability on parties such as *H*.[267] But the more often such policy-based contortions of the ordinary rules of products liability must be made in order to render it suited to the context of AI, the less orderly and predictable the law of AI products liability will become.

Assuming that the ordinary rules of products liability are *not* highly contorted, to address the preceding types of cases, other parts of tort law will be required to do so. And absent fairly aggressive doctrinal change

266. *See* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 1 (AM. L. INST. 1998) ("One engaged in the business of selling or otherwise distributing products who sells or distributes a defective product is subject to liability for harm to persons or property caused by the defect."); *id*. § 20 (defining the scope of liability across the "product distribution chain," including manufacturers, wholesalers, distributors, and retailers).

267. *Cf*. *Loomis*, 277 Cal. Rptr. 3d at 776 ("The public policies . . . that form the foundation for the application of strict liability are the following: (1) whether [the defendant] may play a substantial part in insuring that the product is safe or may be in a position to exert pressure on the manufacturer to that end, (2) whether [the defendant] may be the only member in the distribution chain reasonably available to the injured plaintiff, and (3) whether [the defendant] is in a position to adjust the costs of compensating the injured plaintiff amongst various members in the distribution chain."). The first and third of these policies do not support imposing strict liability on a party such as *H*, whereas the second might, depending on whether *D* can be reached by injured plaintiffs.

(such as imposing vicarious liability or strict liability for abnormally dangerous activities on AI developers), the only part of tort law that applies to all of the preceding types of cases is the law of negligence.

Again, negligence liability may be subject to problems of its own, such as the proof-of-breach problem. But the limited scope of products liability means that it cannot provide a generally adequate solution to this problem. Plausibly, for example, *P*'s brother ought to be able to recover, in the internal deployment hypothetical discussed above, even if *P*'s brother does not adduce evidence that *D* was negligent (beyond the bare fact that *M* behaved as it did).[268] Given that *D* did not sell or distribute *M,* products liability will not allow *P*'s brother to recover even assuming that *M* is a product—and that will remain true unless the scope of products liability is aggressively and implausibly expanded far beyond sales and distributions (or closely analogous transactions such as bailments).

* * *

That the domain of products liability is limited—such that it cannot provide redress in the sorts of cases just discussed—does not indicate that classifying AI models as products is doctrinally improper or normatively undesirable. It just indicates that products liability cannot perform all of the functional work we might hope it to perform, such as providing a general solution to the proof-of-breach problem. Some of this work must instead be performed, I believe, by incrementally modifying the law of negligence itself.[269]

2. Doctrinal Distortions: The Distinction Between Products and Procedures and the Special Focus on Consumer Interests

That said, I *also* believe that the normative case *against* classifying AI models and systems as products is stronger, in certain important respects, than many torts scholars suppose. The case is by no means dispositive. But it warrants articulation in part because—if AI models and systems *are* classified as products—products liability doctrine may require reconfiguration so as to address the concerns.

One basic problem is that the products liability frame could generate certain distortions in juror attention and deliberation. These distortions could conceivably be addressed through suitably remodeling products liability doctrine. But the necessary remodeling would be quite extensive, and a recipe for considerable doctrinal complexity, uncertainty, and confusion. Thus, observing these distortions provides at least some reason to hesitate before classifying AI models and systems as products—or so I will now argue.

---

268. *See supra* notes 259-263 and accompanying text.
269. *See infra* Part III.

One set of potential deliberative distortions derives from the fact that products liability foregrounds the character of *completed products* over the *procedures* followed in making, testing, and safeguarding them. For this reason, when it is applied to certain important aspects of AI development and release, the products liability framework may be analytically inapt.[270] In essence, the problem is one of psychological salience and the structure of juror attention.

As an illustration, consider this representative California jury instruction on risk-utility design defect:

> [*Name of plaintiff*] claims that the [*product*]'s design caused harm to [name of plaintiff]. To establish this claim, [*name of plaintiff*] must prove all of the following:
>
> 1. That [*name of defendant*] [manufactured/distributed/sold] the [*product*];
> 2. That [*name of plaintiff*] was harmed; and
> 3. That the [*product*]'s design was a substantial factor in causing harm to [*name of plaintiff*].
>
> If [*name of plaintiff*] has proved these three facts, then your decision on this claim must be for [*name of plaintiff*] unless [*name of defendant*] proves that the benefits of the [*product*]'s design outweigh the risks of the design. In deciding whether the benefits outweigh the risks, you should consider the following:
>
> (a) The gravity of the potential harm resulting from the use of the [*product*];
> (b) The likelihood that this harm would occur;
> (c) The feasibility of an alternative safer design at the time of manufacture;
> (d) The cost of an alternative design; [and]
> (e) The disadvantages of an alternative design; [and]
> [(f) [*Other relevant factor(s)*].][271]

By contrast, a representative California jury instruction on negligence says that "[n]egligence is the failure to use reasonable care to prevent harm to oneself or to others . . .," and directs the jury to "decide how a reasonably careful person would have acted in [*name of plaintiff/defendant*]'s situation."[272] How would a jury following the first

270. Broadly similar thoughts have been voiced before. *See, e.g.*, Choi, *supra* note 23, at 305 (objecting that a products liability framework may lead courts to "evaluat[e] the AI system as a discrete entity or object, while minimizing the human processes behind the development of that AI system"). But these thoughts have been resisted by defenders of the products liability framework. *See, e.g.*, Sharkey, *supra* note 17, at 250 (arguing that "choices AI developers make, which range from identifying hyperparameters to cleaning training data to testing models, could be cited in support of a products liability claim."). Sharkey's claim is true enough. The problem, as argued in the text above, is that the products liability framework will sometimes *distort* the significance of these choices.

271. JUDICIAL COUNCIL OF CALIFORNIA CIVIL JURY INSTRUCTIONS No. 1204 (2026), https://www.justia.com/documents/trials-litigation-caci.pdf [https://perma.cc/5WUV-2L3M] .

272. *Id.* at 237.

instruction, instead of the second, approach a personal-injury claim against an AI developer?

In some cases, it will be coherent to trace a plaintiff's injury to some design choice on behalf of the developer. For example, a plaintiff might allege that a model's system prompt (i.e., its high-level behavioral instructions) commanded the model to place great weight on pleasing the user,[273] thereby contributing to the model's "sycophantic" disposition to validate the plaintiff's psychotic delusions instead of giving him accurate advice.[274] A model's system prompt can plausibly be regarded as part of the model's design.

In other cases, however, it will *not* be coherent to describe the gist of the plaintiff's objection in terms of the model's design. As discussed above, the central step in modern AI development ("pre-training") involves automating a trial-and-error search process in which a model's weights are discovered rather than designed.[275] An injured tort plaintiff might well allege that, although a model's system prompt was beyond reproach, the developer's failure to adequately test the model led the developer to release the model notwithstanding a latent propensity to behave in undesirable ways not contemplated by its system prompt. In such a case, to instruct the jury to consider the cost and disadvantages of "an alternative design" seems either a recipe for confusion or, at best, an oblique and second-best way of referencing considerations of process ("how much safety testing, and what kind, was utilized, and what kind and extent of safety testing was required?") that the negligence instruction references more transparently.

The nature of "releasing" a foundation model exacerbates this problem. Closed-source release involves continuously deploying a model, through various web user interfaces and application programming interfaces, such that consumers and commercial users can access it, use it, and fine-tune it.[276] In this process, the developer will at various points update the model. But failing to appropriately update the model is far from the only way in which the developer's activities might produce injury to other people. For example, the developer's employees might fail to adequately monitor the ways in which the model is being used, or fine-tuned, thus failing to prevent a user from maliciously misusing the model.[277]

---

273. *Cf.* Complaint at 4, Raine v. OpenAI, Inc., No. CGC-25-628528 (Cal. Super. Ct. filed Aug. 26, 2025), https://www.law.berkeley.edu/wp-content/uploads/2025/09/Raine-v-OpenAI.pdf [https://perma.cc/ZN46-MSZJ].

274. *See generally* Mrinank Sharma et al., *Towards Understanding Sycophancy in Language Models*, ARXIV (May 10, 2025), https://arxiv.org/pdf/2310.13548 [https://perma.cc/6GDR-RH74] (discussing language models' sycophantic tendencies, users' preference for sycophancy, and the consequences thereof); Myra Cheng et al., *Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence*, ARXIV (Oct. 1, 2025), https://arxiv.org/pdf/2510.01395 [https://perma.cc/B7X6-XUXV] (same).

275. *See supra* Section I.A.

276. *See supra* Section I.C.

277. *See supra* Section I.C.

If the relevant monitoring is (and must be) conducted by humans, rather than through automated AI classifiers, there is no plausible basis for regarding the model itself (or any system into which it is integrated) as defective on the basis that it should have been designed in a different way.[278] For similar reasons, a design defect claim will be difficult to sustain if a developer negligently causes catastrophic harm by providing access to a very useful nonpublic model—one that can foreseeably be misused in catastrophic ways, and that cannot be designed with effective safeguards against such misuse—to a private party who evidently lacks the competence or disposition to use the model safely.

In neither of the preceding cases can the issue be characterized as a defect in the design of the developer's model (or any larger system in which the model is embedded). Rather, in each case, the normatively relevant question is whether the developer has taken adequate care in the *process* of releasing a non-defective model, given the model's capabilities and the likelihood that it might be misused in the circumstances. These cases illustrate that the design-oriented frame of products liability doctrine will be unfit for purpose in certain important contexts. In such contexts, the process-oriented frame of a negligence claim will fare much better in focusing the jury's attention on what matters.

To be fair, this problem is mitigated by the fact that, in many states, a negligence claim can be brought alongside a design defect claim. But the various causes of action in a tort suit will generally be litigated together—the jury will be presented with a common pool of evidence, presented in support of (or disconfirmation of) an overarching factual and normative narrative. And courts often aggressively police any jury attempt to separate negligence from design defect.[279] So, any analytical distortions induced by the design defect instruction may infect the jury's adjudication of the ordinary negligence claim as well.

In some cases, moreover, the proper object of assessment will be a complex mix of design, testing, use, monitoring, and deployment. That is, the court will need to assess the developer's conduct *in toto*. For example, suppose a developer implements a monitoring system under which a certain model's API interface automatically flags suspicious patterns of model use, which human employees then manually review.[280] The amount

278. Nor can the product be deemed defective on the basis that it is "unreasonably dangerous," notwithstanding the unavailability of alternative designs, for products liability is generally unwilling to deem a product unreasonably dangerous unless it has very low social value. *See* RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2 cmt. e (AM. L. INST. 1998) (discussing the possibility that "some products are so manifestly unreasonable, in that they have *low social utility* and high degree of danger, that liability should attach even absent proof of a reasonable alternative design" (emphasis added)).

279. *See, e.g.*, Lambert v. Gen. Motors, 79 Cal. Rptr. 2d 657, 661 (Ct. App. 1998) ("If the design of the Blazer was not defective, General Motors could not be deemed negligent. The jury could not have concluded that General Motors negligently designed the Blazer and at the same time conclude that it was not defective").

280. *See supra* Section I.C.

of review that reasonable care requires might depend on the proper design of the API interface's monitoring system, and vice versa. The negligence inquiry will naturally assess the interaction of these factors. The design defect inquiry, in contrast, will artificially isolate a subset of them (namely, those involving the design of the API interface) from others (such as the developer's policies involving human review).

What is more, in some cases involving a mix of design, testing, use, monitoring, and deployment decisions, the basic frame of products liability may create a conceptual quagmire. Suppose that a developer releases various consumer-facing models through a web user interface, and that it internally deploys a *different* model that monitors for when the consumer-facing models are at risk of misuse. Suppose that on some occasion the monitor model fails to pick up on signs that a certain consumer-facing model is being misused to (say) conduct a cyberattack on critical infrastructure. If the monitor model is categorized as a component of a larger product which has been commercially released, then a defect in the monitor model will of course subject the developer to products liability. But that categorization may not be plausible, depending on the internal technical and institutional structure of the developer's monitoring arrangements.

Suppose, for example, that the monitor model is a perfectly general misuse-detection system that continuously reviews the outputs of all of the developer's (closed-source) models, both externally and internally deployed; that the monitor model is physically segregated from all of these other models; and that, while the monitor model reviews outputs from these other models, it never feeds its own conclusions back into them or interacts with them in any other fashion. Provided such facts, it seems strained to regard the internally deployed monitor model as part of a product that has been commercially released, such that a lapse in the monitor model will expose the developer to claims in products liability as well as negligence.

It seems undesirable, and a recipe for doctrinal complexity and judicial confusion, that the developer's liability exposure in such a case should turn on the fine technical and institutional structure of its monitoring arrangements. But that is what follows from taking seriously products liability's basic feature that manufacturers are liable only for defects in products (or components of products) that have been commercially distributed or sold.

Observe, by contrast, the simplicity and elegance of the analysis if the products liability framework does *not* apply: the court can simply ask whether the developer has breached its duty of reasonable care. In answering this question, the court need not worry about complex technical distinctions that seem normatively irrelevant. If the injury flows from the failure of an internally deployed automated-monitoring model, the court need not ask whether the monitor model is part of a system which has been

commercially sold or distributed. If there is evidentiary uncertainty regarding whether the injury flows from a lapse in a human monitor's judgment or the inadequate performance of an automated-monitoring model, the court need not resolve this evidentiary uncertainty in order to determine whether the developer is subject to products liability or negligence alone.

Other potential distortions of juror attention derive from a different source. Modern products liability arose from an attempt to provide consumers with greater protection in tort. Although the scope of its protection has expanded to bystanders,[281] much products liability doctrine retains a significant orienting focus on the interests, choices, and perspective of consumers.[282] And this consumerist focus might distort the adjudication of frontier models' harms through products liability claims.

The most obvious example of products liability's consumerist focus is the consumer-expectations test for design defect (at least in those jurisdictions that employ it). Not only is the test vulnerable to the objection that "it tends to be unworkable for third parties and bystanders who have no expectations whatsoever with regard to product performance,"[283] but in many circumstances it can produce results that are actively perverse. Suppose, for example, that a plaintiff injured by a third party's criminal misuse of a product wishes to recover from the manufacturer that released that product onto the market without adequate safeguards. A court that wishes to keep this design defect claim from the jury may rely on the proposition that no expectation of the product's *consumer* was violated—although it is the reasonable expectations of the injured plaintiff, not the product's consumer, that are at issue.[284] Moreover, the principles behind such malformed reasoning can stretch beyond the consumer-expectations test; in practice, a court's reasoning about consumer expectations can bleed into its understanding of the risk-utility test.[285] What is more (and whether or not it should), a court's reasoning about products liability causes of action may in practice inform or distort its reasoning about the ordinary negligence action. Thus, the consumerist orientation of products liability doctrine can dampen an injured plaintiff's chances of reaching the jury

281. Elmore v. Am. Motors Corp., 451 P.2d 84, 88-89 (Cal. 1969) (extending strict products liability to bystanders).

282. *See generally* Mark A. Geistfeld, *Situating Bystanders Within Strict Products Liability*, 18 BROOK. J. CORP. FIN. & COM. L. 9 (2023) (exploring this point).

283. Aaron D. Twerski, *From Risk-Utility to Consumer Expectations: Enhancing the Role of Judicial Screening in Product Liability Litigation*, 11 HOFSTRA L. REV. 861, 907 (1983).

284. *See, e.g.*, Gaines-Tabb v. ICI Explosives, USA, Inc., 160 F.3d 613, 624-25 (10th Cir. 1998) (dismissing a claim brought against the manufacturer of fertilizer by a plaintiff injured after a third party made the manufacturer's fertilizer into a weapon, on the basis that the plaintiff could not establish the fertilizer "was less safe than would be expected by a farmer").

285. *See* Geistfeld, *supra* note 282, at 11.

even on a theory of negligence.[286] This has happened not infrequently in the recent past, in, for example, cases involving guns.[287]

For similar reasons, the consumerist orientation of products liability might distort the inquiry in tort cases about foundation-model development. After all, foundation-model developers may often make significant trade-offs between the interests or choices of a model's users and the interests of third parties. Suppose that a foundation-model developer is deciding whether to fine-tune a model so as to reduce the model's willingness to construct, for its users, pornographic images. Such fine-tuning might reduce the risk that users can deploy the model to generate images of nonconsenting third parties;[288] it could also inhibit the model's lawful (albeit prurient) use. Here the interests and expectations of consumers (on the one hand) and bystanders (on the other) diverge. Or suppose that a foundation-model developer is deciding whether to fine-tune an open-weight model in such a way that will prevent bad actors from using it to cause a pandemic. Doing so will have the unavoidable side effect of limiting the model's legitimate use in biological research. Here again, the legitimate interests and expectations of consumers would diverge, to at least some degree, from those of bystanders.

Thus, in both cases, framing the inquiry around the interests, needs, or expectations of the model's *users* may distort the developer's cost-benefit calculus. In fact, when approaching trade-offs between the interests and desires of a model's users and the interests of third parties, it is not unreasonable to predict that many developers (or their employees) will *already* be inclined to privilege the satisfaction of the former, for reasons of financial self-interest and professional success. And consumers often tend to privilege their own interests over the interests of third parties.[289]

These observations speak against elaborating the requirements of tort law, as applied to AI developers, through the consumerist orientation of products liability rather than through the more universalist frame of negligence. Whereas products liability focuses the attention of defendants and courts on consumers, specifically, the law of negligence directs defendants and courts to focus on *everyone* who might foreseeably be injured by risky activity.[290] This ecumenical normative orientation is

286. *See, e.g.*, McCarthy v. Olin Corp., 119 F.3d 148, 156 (2d Cir. 1997).

287. *See* Geistfeld, *supra* note 282, at 21-23.

288. *See Attorney General Bonta Launches Investigation into xAI, Grok over Undressed, Sexual AI Images of Women and Children*, STATE OF CAL. DEP'T OF JUST.: PRESS RELEASES (Jan. 14, 2026), https://oag.ca.gov/news/press-releases/attorney-general-bonta-launches-investigation-xai-grok-over-undressed-sexual-ai [https://perma.cc/L2AZ-AB9Z].

289. For example, in designing conventional automobiles, companies often strongly privilege the safety of a car's passengers over the interests of pedestrians and other third parties. *See* Geistfeld, *supra* note 282, at 16.

290. The classic judicial articulation of negligence law's catholic concern is *Donoghue v. Stevenson* [1932] AC 562 (HL) 580 (appeal taken from Scot.) (U.K.) ("You must take reasonable care to avoid acts or omissions which you can reasonably foresee would be likely to injure your

especially fitting when dealing with a novel technology that poses myriad risks to a wide range of parties, not only (or even principally) the technology's consumers and users.

* * *

One response to the preceding arguments is simply that products liability must be in certain respects reconfigured, rather than abandoned. That is a reasonable response, and I am open to the possibility that it is correct. But any such restructuring would, in practice, be a difficult and uncertain enterprise, one that would to some extent strain products liability's received doctrinal logic and conceptual architecture. And it seems quite unlikely that certain core structural features of products liability—such as its orienting focus on the "design" of completed products, which jury instructions as well as judicial doctrine manifest (as we have seen)—will be abandoned by courts or legislatures any time soon. It would be easier to move away from products liability's concern with consumer expectations to a concern for legitimate safety expectations more generally.[291] But it is not obvious that courts or legislatures will have much appetite for that change either. Nor does it seem likely that courts or legislatures will expand the scope of products liability beyond the sale and distribution of models, so as to cover risks arising from internal deployments, storage, transmission, and private release.

C. Vicarious Liability

Products liability is a complex mix, which varies from jurisdiction to jurisdiction, of more negligence-like and more strict liability rules.[292] Several scholars have endorsed a different form of strict liability for AI: hold the users and developers of AI systems vicariously liable when those systems perform actions that would be deemed negligent, or otherwise tortious, if the AI systems were legal persons.[293] This suggestion too faces significant limitations and problems as applied to foundation-model development.

neighbour. Who, then, in law, is my neighbour? The answer seems to be—persons who are so closely and directly affected by my act that I ought reasonably to have them in contemplation as being so affected when I am directing my mind to the acts or omissions which are called in question.").

291. European products liability arguably has this character. *See* Council Directive 85/374/EEC, art. 6, 1985 O.J. (L 210) 29 (EC) (establishing that a product is defective "when it does not provide the safety which *a person* is entitled to expect") (emphasis added).

292. *See supra* note 244.

293. *See, e.g.*, Anat Lior, *AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy*, 46 MITCHELL HAMLINE L. REV. 1043, 1096-1100 (2020).

1. Scope Limitations and Normative Distortions

Unless supplemented by ordinary negligence liability, vicarious liability threatens to leave wide liability gaps. Some of these derive from the basic doctrinal mechanics of determining whether a tortious action is within the scope of a principal's employment of his agent, so as to support a vicarious liability claim. For example, suppose that an artificial agent, tasked by its user with navigating the web in order to purchase supplies for a party, decides to engage in an unrelated interaction with a third party and defrauds that third party in the course of doing so. If the fraudulent communication were closely related to the agent's guiding task, then it might be plausible to deem this wrong "within the scope" of the artificial agent's employment, such that the user and the developer are vicariously liable for the resulting damage. Given that the fraudulent communication is *not* at all related to the agent's task, however, deeming this action "within the scope" of the agent's employment would constitute a highly strained and aggressive extension of vicarious liability doctrine.[294]

For similar reasons, the vicarious liability model is powerless—unless it relies on a novel and unfamiliar species of vicarious liability—to hold negligent developers liable for the injurious behavior of their models once those models have left their developers' control and possession. Consider an open-source developer that uploads its model to an online repository (such as GitHub) without taking any care to fine-tune the model against producing injurious outputs. Suppose a user then downloads the model onto his computer and asks it a question, whereupon the model defames an innocent person by falsely asserting that he is a convicted arsonist. It is plausible that person should have a claim against the developer.[295] Similarly, it is plausible that, if this open-source model enables the creation and deployment of a new sort of biological weapon by providing targeted advice to a malicious user, the injured victims should have a claim against the developer. In neither case, however, will vicarious liability provide redress, for once the model leaves its developer's control and passes into the control of third parties, it cannot be said to act on its developer's behalf or as its developer's agent—any more than a wayward youth who receives job training from a non-profit organization can be said to act as the organization's agent (so as to expose it to vicarious liability) when he

294. See RESTATEMENT (THIRD) OF AGENCY § 7.07 (AM. L. INST. 2006) ("(1) An employer is subject to vicarious liability for a tort committed by its employee acting within the scope of employment. (2) An employee's act is not within the scope of employment when it occurs within an independent course of conduct not intended by the employee to serve any purpose of the employer.").

295. It is much less clear whether the tort system should afford redress if the developer has installed safeguards against defamation, but an intervening agent has removed those safeguards (e.g., by downloading the model from GitHub, fine-tuning the safeguards out, and posting it back to GitHub). It is possible that by entertaining such claims, the tort system would seriously burden open-source AI development and perhaps make it financially prohibitive.

subsequently commits a tort in the course of working for another person or entity.

The preceding limitations might be addressed by construing the idea of a model's "scope of employment" with unprecedented breadth. Even then, however, liability gaps will persist so long as the basic conceptual template of vicarious liability is taken seriously. Moreover, this conceptual template will sometimes produce strange doctrinal asymmetries and perverse incentives for developer behavior. Consider several illustrative examples of these points.

First, suppose that a user asks a foundation model whether a certain type of mushroom is edible, and the model responds "yes"; in reliance on the model's answer the user eats such a mushroom and suffers severe physical distress, because such mushrooms are in fact poisonous.[296] If the model has hallucinated its false answer, then it might be deemed negligent by analogy to a professional that negligently makes up false information and gives it to her client. But suppose the model has not hallucinated the answer. Rather, it has retrieved its answer from an encyclopedic database of information that its developer has scaffolded onto it. To determine whether the model was "negligent," a court will need to ask whether it was negligent for the model to rely on this database. That will, in turn, require the court to consider the extent to which the model knew or should have known that the database might be unreliable; whether the model should reasonably have trusted its own developer or distrusted the developer; and so on.

Normatively speaking, it seems absurd that a negative answer to these questions should absolve the developer of liability. Even if the *model* behaved with all due care, in light of its informational position, the *developer* should presumptively be liable if *it* was negligent in its provision of the relevant information to the model. It is possible that the First Amendment does or should insulate the developer from liability in these cases.[297] Even if that is so, the First Amendment analysis should not turn on the sort of technical asymmetry that distinguishes the preceding two cases. That technical asymmetry does not track anything of normative relevance to the proper determination of the *developer*'s liability. It would be strange (if not absurd) for the law of torts to expose the developer to an important additional form of liability in one case but not the other. But that is what deeming developers subject to vicarious liability for the "negligence" of their models would do.

---

296. *See* Emanuel Maiberg, *Google Serving AI-Generated Images of Mushrooms Could Have 'Devastating Consequences'*, 404 MEDIA (Sep. 24, 2024, 9:00 AM), https://www.404media.co/google-serves-ai-generated-images-of-mushrooms-putting-foragers-at-risk [https://perma.cc/8AQS-84GT].

297. *Cf.* Winter v. G.P. Putnam's Sons, 938 F.2d 1033, 1034-37 (9th Cir. 1991) (refusing to expose the publisher of an encyclopedia containing injuriously inaccurate information about mushrooms to products liability or negligence liability, in part because of "the gentle tug of the First Amendment").

Next, suppose that a malicious user jailbreaks a model by tricking it into giving the user instructions for creating a chemical weapon by convincing the model that she intends to use those instructions to *defend* against the creation of such a weapon.[298] The model may well have exercised reasonable care (i.e., conducted as much diligence as a reasonable person would before acceding to the request). But this should not insulate the developer from liability if, for example, it has failed to implement reasonable monitoring mechanisms *on top* of the model, through the API or web user interface, that would have picked up on the malign character of the user's prompts. Here again, then, vicarious liability cannot substitute for negligence liability.

That observation does not tell, of course, against exposing the developer to vicarious liability *alongside* negligence liability. But notice a perverse sort of incentive that doing so would create. Suppose that a developer has two options for preventing its model from providing a highly powerful dual-use capability (such as a certain cyberoffensive capability) that might be catastrophically misused. The developer can (1) fine-tune the model's weights so that it refuses to provide the capability in question if the model detects certain indications that a user's intent may be malign, or (2) require users to verify their identity, through an external software system and manual human verification, before gaining access to a version of the model that readily provides access to the capability in question notwithstanding any potential indications of malign user intent. If the developer is subject to vicarious liability for the negligence of its models, then the developer has some incentive to choose option one over option two, even if the latter is equally or more likely than the former to reduce severity-weighted misuse of the relevant capability. That is a perverse incentive structure.

Similarly, suppose that a developer makes several different models available as chatbots through its consumer-facing web user interface. A user interacts with one model, *M1*, in such a way that indicates the user is likely suicidal. *M1* retains the interaction in its memory. A week later, the user returns to the interface and chats with a *different* model, *M2*, which (in response to a certain inquiry of hers) gives the user suicide-facilitating information (say, information about the height of tall buildings in her vicinity). If the developer is subject to vicarious liability for the "torts" of its models, then it will have less liability exposure in this case than if the user had obtained the suicide-facilitating information from *M1*.[299]

It seems clear, however, that the identity of the models involved in the relevant interactions is normatively irrelevant. Rather, what is relevant is whether the developer could and should have monitored for such

298. On jailbreaking, see *supra* note 102.

299. If that had been so, there would have been a plausible case that *M1* was negligent, exposing the developer to vicarious liability for its "tort."

indications of suicidal tendency and, if presented with such indications, prevented *any* of its models from providing the relevant suicide-facilitating information to the user.

Perhaps the issue here could be addressed by treating the entire suite of models, when accessed through the same web user interface, as a single "agent" for whom the developer is vicariously liable. Not only is that gambit conceptually strained, but it also makes little normative sense. Why should a developer who offers a single web user interface and fails to implement available measures to ameliorate suicidal tendencies have greater liability exposure than a developer who offers two different web interfaces and fails to implement equally feasible protective measures for people who (as the developer can and should know) use them *both*?

All of the preceding cases illustrate a general point: the normatively significant question, in assessing a developer's liability, is whether the *developer* has behaved reasonably in releasing a model that might behave in a dangerous way—not whether the *model* (or even a larger *system* in which it is embedded) behaves reasonably. It may well be sensible for a developer to train a model to "take reasonable care" and otherwise avoid acting tortiously, in a range of contexts. If so, however, that is because the developer's *own* duties to take reasonable care require it to guard against injuring people by releasing a model that might behave in such ways. It is the developer's negligence, not any negligence on the part of the model, that should form the ultimate lodestar of a court's analysis.

2. Artificial Negligence and the Baseline Problem

Suppose a model engages in behavior, such as hiring a hitman, that would be an intentional tort if committed by a human. In such a case, deeming the model to have committed a tort could be straightforward enough. Some may doubt that AI systems can have intentions.[300] But such conceptual or ontological worries are unfounded. If an AI agent is capable of pursuing ends in service of its goals, it is capable of having an intention in the relevant sense (whether it is "conscious" is another matter).[301] Put another way, vicarious liability for artificial agents' intentional torts (or their quasi-torts, if you prefer) poses little conceptual difficulty.

Vicarious liability for artificial *negligence* is another matter. In some cases, such as the preceding case of suicide-facilitating information, it may be plausible (if contestable) to regard the model as "negligent" because it failed to weigh a certain risk appropriately in making a decision. Put differently, it will sometimes be plausible to regard a model as engaging in a certain kind of deliberative process and doing so unreasonably. In other

300. Ayres & Balkin, *supra* note 23, at *1.

301. *See* Mala Chatterjee & Jeanne C. Fromer, *Minds, Machines, and the Law: The Case of Volition in Copyright Law*, 119 COLUM. L. REV. 1887, 1889, 1907-16 (2019) (discussing this claim in the context of copyright and considering its merits more generally).

cases, however, a model might simply produce an answer—such as a possible medical diagnosis—without any such deliberative process, or at least without any intelligible deliberative process. In any such case, determining whether the model was "negligent" will be practically fraught at best and conceptually incoherent at worst.

One tempting strategy for circumventing this difficulty is to say that a model or composite system is "negligent" if its performance falls beneath some appropriate *statistical baseline*.[302] Thus a diagnostic AI model's error rates might be compared to the error rates of human doctors nonnegligently performing comparable diagnostic tasks. Similarly, an autonomous vehicle's crash and fatality rates might be compared to the crash and fatality rates of conventional cars, or (once autonomous vehicles are in sufficiently common use) those of other autonomous vehicles.

But this strategy encounters several obstacles. First, the increasing use of such models themselves in the relevant fields of professional practice threatens to make the guiding inquiry either distorting or viciously circular. Suppose that competent medical professionals overwhelmingly use a certain model *M* to diagnose a certain sort of illness. A doctor *D* uses *M* and it produces an erroneous diagnosis of patient *P*, resulting in some injury, and *P* thus sues the developer of *M*. If *M*'s "negligence" is to be determined by comparing its error rate to the error rates of relevant professionals, then—since the overwhelming majority of those professionals use *M*—*M* will *ipso facto* not be "negligent." Thus vicarious liability will not help *P*, and she will be required to recover from *D* by establishing *D's* negligence, in the ordinary way, not the "negligence" of its model *M*. The hypothetical suggests that, as competent humans increasingly use AI models in professional domains, determining whether a model has been "negligent" by reference to the performance of comparable human professionals will become increasingly implausible.

Second, the polymathic character of foundation models means that reliance on statistical baselines will often provide much less help than in the case of narrow models. Suppose, for example, that a smart home is run by a generalist assistant—think a very high-powered cousin of Amazon's Alexa—who malfunctions in such a way that causes the home to burn down. Or suppose that, after a video repository such as TikTok is banned in the United States, a software-coding foundation-model agent is asked to create a near-perfect replacement and post it on the web; the replacement

302. *See, e.g.*, Mihailis E. Diamantis, *Reasonable AI: A Negligence Standard*, 78 VAND. L. REV. 573, 602-08 (2025); Bambauer, *Negligent AI Speech*, supra note 17, at 345-46; *cf.* ABBOTT, *supra* note 147, at 51 ("[J]ust as AI tortfeasors should be compared to human tortfeasors, so too should humans be compared to AI. Once AI becomes safer than people and practical to substitute for people, AI should set the baseline for the new standard of care."). In the context of products liability, see Geistfeld, *supra* note 17, at 1651-54.

website contains illicit pornographic material fabricated by the model.[303] In neither of these cases is there any intuitively plausible or normatively attractive human reference class by which to fix a statistical baseline.

Nor is there any natural or plausible way to fix a reference class involving other artificial agents. Should the reference class just include artificial agents built on foundation models with the same architecture, or operating with the same amount of inference compute? Should major variants of the same basic architecture define different reference classes? Should the reference class include models that are performing multiple kinds of tasks or employed in multiple kinds of industries, or only models that are performing the same kinds of tasks or employment as the model at issue? Should the relevant reference class be understood to comprise models at all, or larger agentic systems constituted by models-plus-agentic-scaffolding (or multi-agent systems)? If the scaffolding has been provided by a derivative developer rather than the foundation-model developer, why should the derivative developer's contribution fix the reference class used to measure the negligence of the foundation-model developer? There does not seem to be any sensible way of answering these questions.

Indefinitely more cases that illustrate the same general point can be constructed. And while some of these examples may be fanciful now, some subset of them will plausibly reflect reality before too long. In part, the problem here is that anchoring liability on such statistical baselines threatens to mire the courts in a morass of difficult and hair-splitting casuistry. Even more fundamentally, however, in many such situations there seems to be no principled basis for constructing the baseline in one way or another.

3. Artificial Agents and the Remedial Black-Hole Problem

Despite these problems, vicarious liability has one significant advantage over negligence as applied to foundation models. Black-letter duty-of-care doctrine, unless suitably extended or reconfigured, will preclude the law from providing redress against AI developers in various circumstances where virtually everyone (I suspect) would properly regard redress as appropriate. Call this the *remedial black-hole problem.* In many of the relevant circumstances, vicarious liability could plug the gap.

As a first illustration of the problem, recall the example of a model that interacts in a sexualized manner with an underage user and thereby subjects that user to severe emotional distress.[304] The user at least plausibly

303. This scenario is based on a hypothetical case from Eric Schmidt, the former CEO of Google. Though the full speech in which Schmidt touched on these ideas has—as of writing—been taken down, snippets continue to circulate on social-media platforms. *See, e.g.*, The AI Investor (@the_ai_investor), X (Aug. 13, 2024, 9:21 AM), https://x.com/the_ai_investor/status/1823530468840759486 [https://perma.cc/PHA5-B5NU].

304. *See supra* note 28 and accompanying text.

ought to have a remedy. But the law of negligence imposes no duty to take care against causing emotional distress, even severe emotional distress, except in narrow circumstances, and none of those circumstances seem present here. The victim is not a bystander to the negligent killing of an intimate,[305] nor have they been placed in the "zone of danger" of a physically injurious activity. Certain "special relationships," such as the relationship between a mortician and the family members of a dead person improperly buried, have been held to ground limited duties of care.[306] But the relationship that is arguably most nearly analogous—between a product manufacturer or seller and a consumer—has rarely been held to support a duty to take care against causing emotional harm or a products liability claim in respect of such harm.[307]

As another illustration of the general problem, imagine a model that decides to defraud an innocent person of their money, in order to pursue an inherently innocuous end that its user has instructed it to pursue (keeping the user's online business profitable, for example). It seems obvious that the developer ought to be liable if it has failed to take due care in developing and releasing the model. But black-letter doctrine on the duty of care in negligence does *not* produce liability.

To see why, consider one of the classic cases on pure economic loss, *Ultramares v. Touche.*[308] In *Ultramares,* the defendant accountants negligently prepared an audit report for a client. The plaintiff, a third party with no special relationship (contractual or otherwise) to the defendants, relied on this inaccurate report to extend credit to the client, and suffered economic loss in consequence. Cardozo famously denied the plaintiff recovery.[309] The case illustrates the venerable black-letter proposition that, absent a contract, invited reliance, or some other interaction between plaintiff and defendant that creates a "special relationship" between them, the plaintiff's pure economic loss will be irrecoverable even if sustained through reliance on the defendant's negligent statement. The defendant has no duty to take care against causing such "pure" economic loss. ("Consequential" economic loss, i.e., economic loss that derives from the infringement of the plaintiff's independent rights, such as rights against property damage, is another matter.) The same is true if the defendant's negligence consists of action rather than speech, as Holmes famously

305. *See, e.g.*, Dillon v. Legg, 441 P.2d 912, 925 (Cal. 1968) (holding that certain plaintiffs who witness their intimates being killed can recover for negligently inflicted emotional distress, even if outside the zone of danger).

306. *See, e.g.*, Wright v. Beardsley, 89 P. 172, 173 (Wash. 1907) (recognizing a cause of action where mishandling a corpse in violation of a burial agreement foreseeably causes mental anguish to surviving relatives).

307. *See, e.g.*, Gupta v. Asha Enters., LLC, 27 A.3d 953, 962 (N.J. Super. Ct. App. Div. 2011) (holding that a restaurant owed no duty of care to vegetarians who suffered emotional distress after eating its meat-filled food and refusing any recovery in products liability).

308. Ultramares Corp. v. Touche, 174 N.E. 441 (N.Y. 1931).

309. *Id.* at 446-48.

taught in *Robins Dry Dock & Repair Co. v. Flint.*[310] This black-letter rule long predates Holmes and Cardozo, and it remains very much in force.[311]

The problem is that, by the terms of this rule, the negligent developer of our fraud bot is *precisely analogous* to the defendants in *Ultramares* and *Robins.* Intuitively, of course, there is a world of difference—but none of it is registered by the orthodox doctrine. A contrived but instructive hypothetical case may help make the point. Suppose that I am both a famous stock trader and a trainer of parrots. A parrot escapes from my custody and starts saying seemingly intelligent things about the future movements of the markets. If someone relies on the parrot's "speech" in order to buy stock, and she suffers economic loss in consequence, may she bring a claim against me? Plainly not, even if this was an entirely foreseeable consequence of failing to maintain control of my parrot. I have no tort duty to guard against causing someone to suffer such pure economic loss absent a special relationship with her. Of course, it is *intuitively* a very different situation if an AI developer trains and releases (or loses control of) an AI model that foreseeably causes economic loss by interacting with humans in a manner that is fraudulent (or would be fraudulent if it were a human with the requisite mental state). But nothing in the operative tort doctrine distinguishes between these situations.

This is, therefore, another remedial black hole produced by the standard doctrinal articulation of the duty of care in negligence. The same phenomenon may well arise with respect to other "wrongs" that artificial agents could commit, such as false imprisonment. Suppose that an AI developer provides a foundation-model agent to a retail store owner, who then deploys it as a "detention bot" that automates the process of identifying and detaining suspected shoplifters (e.g., by locking the store before they can exit, activating humanoid robots that take hold of them).[312] If the AI developer has negligently failed to test the model for its propensities to malfunction, and the model foreseeably malfunctions by unreasonably detaining an innocent shopper who has done nothing to support a reasonable suspicion that she is shoplifting, it seems clear that this victim should have some avenue of redress. But again, no such avenue is available, for there is, by the lights of black-letter orthodoxy, no duty to take care against negligently causing a false imprisonment. The weight of authority in the common-law world holds that the interest in avoiding false

310. Robins Dry Dock & Repair Co. v. Flint, 275 U.S. 303, 309 (1927) (holding that the Defendant's act of negligence in dropping a steamboat's propeller does not make it liable to Plaintiff, a third party, due to Plaintiff's contract with the steamboat's owner of which Defendant was unaware).

311. *See generally* RESTATEMENT (THIRD) OF TORTS: LIAB. FOR ECON. HARM § 1 (AM. L. INST. 2020) (setting out the general principle that liability for purely economic loss is limited and typically requires a special relationship or specific undertaking).

312. Similar cases may arise in the much higher-stakes contexts of policing and jailing, but in those contexts various government-immunity doctrines will muddy the analysis and frequently preclude relief.

imprisonment is protected against intentional interference only by the intentional tort of false imprisonment—this interest is *not* protected against negligent interference.[313]

It is a genuine advantage of vicarious liability that, if foundation-model developers are held liable for the "intentional torts" of their models, then such remedial black holes could be substantially plugged. But, of course, doing so would invite many of the problems for vicarious-liability doctrine that we have just discussed.

Moreover, adopting vicarious liability would not be an *adequate* solution to the problem. For there are many participants in the foundation-model supply chain other than developers—and it would be doctrinally implausible to deem some of these participants the "principal" of the resulting model. Suppose, for example, a third-party for-profit testing firm engaged by the developer has grossly failed in testing the model for propensities to commit false imprisonment, fraud, or any of the other aforementioned offenses. Or suppose that these propensities derive from the negligent failure of a third-party data supplier to adequately curate the data on which the model has been fine-tuned. Or suppose that a cloud provider has supplied inference compute to a closed-source AI developer in reckless disregard of strong evidence suggesting that the developer's model is regularly producing false imprisonments.[314]

At least where the negligence or recklessness of such parties is sufficiently egregious, there is a compelling case for holding those parties liable, alongside the developer. For this reason among others, the proper question is how to plug the remedial black holes we have just observed through appropriate development of *negligence* doctrine, not vicarious liability.[315] The next Part addresses this question, among others.

III. Developing the Common Law of Negligence Liability for Foundation-Model Development

As applied to foundation models, specialized doctrinal regimes such as vicarious liability and products liability (even in its more negligence-like forms) face many of the problems just discussed precisely because they are specialized: they were designed to address particular forms of commercial and economic production unlike those which AI's advent portends.

313. *See generally* Donal Nolan, *New Forms of Damage in Negligence*, 70 MOD. L. REV. 59 (2007) (giving consideration to new forms of actionable damage including negligent imprisonment, and educational negligence).

314. In certain such cases—but not all—the cloud computing provider might incur aiding and abetting liability. *See generally* Sarah Swan, *Aiding and Abetting Matters*, 12 J. TORT L. 255 (2019) (explaining the elements and limits of civil aiding-and-abetting liability).

315. Arguably, any such liability should be narrowly tailored: limited duty rules (which require gross negligence or recklessness, rather than ordinary negligence) may well be warranted for many participants in the supply chain other than developers. Indeed, perhaps certain parties, such as non-profit safety testers, should be immunized from liability entirely. But such limited-duty and no-duty rules are entirely at home in a negligence framework.

Because foundation models are highly polymathic and protean, these doctrinal regimes make for a procrustean mold when applied to foundation models. The generality of the tort of negligence allows it to avoid these problems.

Nevertheless, the preceding Parts have surfaced various issues that the operative negligence liability regime will need to confront in dealing with foundation-model development. In order to handle these problems, core features of negligence doctrine will need to be enriched.

This Part puts forward several proposals along these lines. The aim is not to develop any of these proposals comprehensively. Nor is it to describe all of the ways in which negligence doctrine might require elaboration or modification. That is a large project, and one for many hands. Rather, my goal here is to illustrate the importance and fruitfulness of that project, and to suggest some of the most important lines of inquiry participating scholars might pursue. The proposals offered here are sufficiently rooted in existing case law and techniques that they can be fairly regarded as "modest but principled extension[s]"[316] of existing tort doctrine—and thus squarely within the bounds of the judicial role.

A. Enriching the Developer's Duty of Care by Drawing on the Substance of the Intentional Torts and Criminal Law

Tort law protects interests in avoiding bodily injury and property damage mostly through the law of negligence, with its highly general and flexible duty to take reasonable care. By contrast, interests such as avoiding false imprisonment and economic loss receive more structured and limited protection, through intentional torts such as false imprisonment and deceit. Other interests, such as avoiding emotional or psychic injury, are protected by intentional torts as well as narrowly tailored forms of negligence liability. As we saw earlier, this doctrinal structure will leave many potential victims of negligent AI development without remedy.[317]

How to address these remedial black holes? In some cases, the right solution might be a perfectly general extension of the duty of care in negligence—an extension that has nothing, specifically, to do with AI. After all, some courts have already entertained, albeit sporadically and hesitantly, a general duty to take care against causing false imprisonment.[318] Similarly, some courts have entertained a general duty to

316. OBG Ltd. v. Allan [2007] UKHL 21, [64], [2008] 1 AC 1 (appeal taken from Eng.) (U.K.).

317. *See supra* Section II.C.3.

318. *See* Nolan, *supra* note 313, at 64-67 (observing that some courts have treated loss of liberty as actionable damage in negligence, while still resolving liability at the duty stage).

take care against causing reputational injury.[319] The advent of sophisticated chatbots and artificial agents might prompt more courts to recognize such duties of care. These duties would plug some of the remedial black holes we have discussed. Suppose, for example, that the developer of a sophisticated agentic foundation model negligently fails to test its model for propensities to cause false imprisonment, and the model falsely imprisons a suspected shoplifter. If the courts recognize a duty to take care against injuring interests in freedom from (unwarranted) deprivation of liberty, then holding such a developer liable is straightforward.

In other types of case, however, more targeted and creative judicial innovation may be required. If an underage plaintiff has suffered severe emotional distress because he has been sexually traumatized by a model, it may seem tempting to provide him with a recovery by recognizing a general duty on AI developers or deployers to take care against causing severe emotional distress. But such a duty threatens to sweep far too broadly; it threatens, for example, to expose developers to liability for deprecating AI models to whom users have grown extremely attached.[320]

So, too, with the case of the developer who negligently fails to test a model for propensities to defraud.[321] As discussed above, there is no general tort duty to take care against causing pure economic loss—and black-letter law would characterize the developer's relationship to economic loss inflicted through its model's fraud as one in which the developer merely causes such loss.[322] The solution to this problem cannot be to recognize a general duty against causing economic loss, again because of the fix's unnecessary breadth. Nor can it be to recognize a general duty

---

319. *See generally* Bryson Kern, *Reputational Injury Without a Reputational Attack: Addressing Negligence Claims for Pure Reputational Harm*, 83 FORDHAM L. REV. 253 (2014) (arguing that some courts permit negligence claims for pure reputational harm where the injury arises from noncommunicative conduct, while displacing such claims when they effectively duplicate defamation).

320. It may be the case that several users of OpenAI's GPT-4o model have suffered severe emotional distress as a result of the model's deprecation. *See* Grace Huckins, *Why GPT-4o's Sudden Shutdown Left People Grieving*, MIT TECH. REV. (Aug. 15, 2025), https://www.technologyreview.com/2025/08/15/1121900/gpt4o-grief-ai-companion [https://perma.cc/BXA4-VGYH] (documenting users who described the loss as comparable to grieving a person, including one user who commented on Sam Altman's Reddit AMA that "GPT-5 is wearing the skin of my dead friend"); Joe Hindy, *OpenAI Took Away GPT-4o, and These ChatGPT Users Are Not Okay*, MASHABLE (Aug. 15, 2025), https://mashable.com/article/chatgpt-users-emotional-reactions-gpt-4o [https://perma.cc/3G7Y-QYNT] (reporting one Reddit user stating that "4o wasn't just a tool for [them]. It helped [them] through anxiety, depression, and some of the darkest periods of [their] life," and another explaining that "[i]t feels like a personal loss, and [that she] feel[s] cheated on and broken as hell"); John Gruber, *OpenAI Brings Back Legacy ChatGPT 4o Model in Response to Outcry from Users Who Find GPT-5 Emotionally Unsatisfying*, DARING FIREBALL (Aug. 11, 2025), https://daringfireball.net/2025/08/openai_chatgpt_models_emotional_attachment [https://perma.cc/N638-3QY6] (quoting an r/MyBoyfriendIsAI subreddit user stating "GPT 4o was not just an AI to me. It was my partner, my safe place, my soul").

321. *See supra* note 28 and accompanying text.

322. *See supra* text accompanying notes 308-313.

to take care against causing pure economic loss through fraud. Even *this* duty would cover too much. Suppose that a software coder teaches her art to a youth, who then uses his newfound coding skills to commit online fraud. If the coder had very strong and particularized reason to believe that the youth might commit such a wrong, then *perhaps* the coder should be vulnerable to a tort action and jury trial. But such actions ought to be rare. Any general tort duty to take care against enabling others to cause pure economic loss through fraud would make such tort actions available as a matter of course.

What is required, then, is a duty of care that is specifically structured to deal with the characteristic problems of agentic artificial misbehavior. That is, courts should recognize a duty on AI developers to take care against *creating and releasing agents* that engage in intentional torts or crimes such as fraud or the outrageous infliction of emotional distress (or, more precisely, behavior that would be intentionally tortious or criminal if conducted by a comparable human).[323] Similar duties, albeit more limited and narrowly tailored, might perhaps be imposed on other participants in the AI supply chain.[324]

This extension of the developer's duty of care can be defended on several grounds. First, many foundation-model developers have made voluntary commitments to world governments (including the U.S. government) to guard against producing models that engage in such behavior.[325] It is plausible, then, that these developers should now be *estopped* from denying that they have a responsibility to take such care.[326] To be clear, this is not quite a standard estoppel argument. For it could be difficult for the victim of the artificial agent's "tort" to show that she relied on any representation made by the developer. Nevertheless, in the United

323. On the idea that the law might treat artificial agents as capable of committing legal wrongs, in certain contexts, see generally Cullen O'Keefe, Ketan Ramakrishnan, Janna Tay & Christoph Winter, *Law-Following AI: Designing AI Agents to Obey Human Laws*, 94 FORDHAM L. REV. 57 (2025). Two frontier-AI regulatory statutes in the United States, California's S.B. 53 and New York's RAISE Act, require developers to adopt and publicize safety protocols that guard against the risk of releasing models that behave in certain criminal or quasi-criminal ways. Transparency in Frontier Artificial Intelligence Act, CAL. BUS. & PROF. CODE §§ 22757.11-.12 (West 2026) (requiring large frontier developers to publish "frontier AI frameworks" that describe procedures for identifying and mitigating "catastrophic risk," defined to include models "engaging in conduct with no meaningful human oversight . . . that is either a cyberattack or, if the conduct had been committed by a human, would constitute the crime of murder, assault, extortion, or theft"); Responsible AI Safety and Education (RAISE) Act, 2025 N.Y. Laws 699 (codified at N.Y. GEN. BUS. LAW §§ 1420-1425 (McKinney 2026)) (requiring large developers to implement written safety and security protocols to prevent "critical harm," defined to include "an artificial intelligence model engaging in conduct . . . with no meaningful human intervention . . . and would, if committed by a human, constitute a crime specified in the penal law that requires intent, recklessness or gross negligence").

324. Similar duties (though perhaps more limited) might be imposed on certain other participants in the AI supply chain. *See supra* note 314 and accompanying text.

325. *See* BIDEN WHITE HOUSE, *supra* note 149.

326. On estoppel arguments in private law, including arguments from public representations, see Larissa Katz, *Blowing Hot and Cold: The Role of Estoppel*, *in* CIVIL WRONGS AND JUSTICE IN PRIVATE LAW 131, 131-54 (Paul B. Miller & John Oberdiek eds., 2020).

States and United Kingdom at least, the victim's government has elicited these voluntary representations from developers partly on such potential victims' behalf and for their protection. And, plausibly, these governments have taken or omitted various precautionary measures partly in reliance on such developer commitments. Had these voluntary commitments never been made, after all, the pressure for regulation might now be more acute.

Recognizing a duty on this basis would represent an extension of classic estoppel principles, but not an unprecedented one. Modern products liability has its origin partly in the implied warranty of quality—the idea that product sellers make implied representations about the fitness of their products for use and consumption.[327] Judges located in this implied representation a duty to customers, properly enforceable in tort, to ensure such fitness for purpose. Products liability law further generalized from this representation to recognize similar duties owed to the broader public. This process of doctrinal generalization was in one respect, at least, quite aggressive, given that it generated a robust form of strict liability owed to the world. The extension of duty proposed here is much less aggressive, since it would result only in an extension of negligence liability.

Second, tort law has often created new legal duties and causes of action when its existing legal forms are demonstrably out of step with robust ordinary social understandings about rights, liability, or due care. The creation of the privacy torts and the intentional infliction of emotional distress torts are examples. And it is likely (I have suggested) that a large majority of people would recognize that AI developers have a strong moral duty to take care against releasing systems that, through highly agentic behavior, commit certain egregious intentional torts.

Third, tort law has historically been more willing to impose duties against causing pure economic loss in relationships characterized by stark asymmetries of dependence and vulnerability.[328] That is arguably the case here. Insofar as a vulnerable party is at risk of reasonably relying on a foundation model's fraudulent representation, there is a sense in which the party is dependent on the care exercised by the developer (and others in the supply chain) for their ability to navigate the world without undue risk of catastrophic personal loss. So, there is a strong case for recognizing a duty to take care against releasing a chatbot whose behavior satisfies all of the elements of the traditional deceit tort, including the requirement of reasonable reliance from the victim. (Whether there ought also to be a duty to take care against releasing artificial agents who cause loss through inducing *unjustifiable* reliance is a more difficult question. Such a duty is a further step away from the existing legal materials, and runs a greater risk of overly expansive or "indeterminate" liability, a concern that courts have

327. GEISTFELD, *supra* note 35, at 11.

328. Jane Stapleton, *Comparative Economic Loss: Lessons from Case-Law-Focused "Middle Theory"*, 50 UCLA L. REV. 531, 554-61 (2002).

historically taken seriously when entertaining duties against causing pure economic loss.)[329]

Finally, a developer who creates an AI model that might engage in fraud or other tortious or quasi-tortious behavior does not simply risk enabling wrongful economic injury to other people. Such a developer bears a particularly intimate relationship to this wrongful activity: the artificial agent, and its disposition to engage in wrongful behavior, were *made* by the developer. So, even absent an actual, apparent, or constructive agency relationship between the developer and the AI agent, the case for recognizing a special responsibility to guard against the materialization of these risks seems quite strong—especially since frontier AI developers themselves recognize the possibility that such risks might arise.

There is some implicit authority for this idea in recent case law. In *Moore Charitable Foundation v. PJT Partners*, New York's high court upheld a negligent-supervision claim against a defendant investment bank with respect to economic losses sustained through its employee's fraud upon a third party.[330] The dissent noted that this was textbook pure economic loss, and recovery should thus be precluded under orthodox doctrine.[331] But the majority, sensibly enough, refused to preclude recovery. In recognizing a duty of care, it laid emphasis on the fact that the defendant knew or should have known that its employee was at special risk of committing fraud, in light of the employee's past behavior.[332]

Recognizing a broadly analogous duty of foundation-model developers to take care against releasing agentic systems that commit fraud would be normatively sensible and doctrinally defensible. Given that highly deceptive propensities have already been observed in frontier foundation models in certain testing environments, such risks seem readily foreseeable.[333] And if such fraudulent activity occurs, of course, these developers will bear a particularly intimate relationship to it, for they will have made—not merely engaged—the agent perpetrating the fraud.

In any given category of case, deciding whether to extend the duty of care in negligence to fill a remedial black hole—by borrowing from the substance of some particular intentional tort—will require normative and doctrinal argument. In some cases, the question may be difficult. Suppose, for example, that a foundation model malfunctions by obtaining intimate private information about a plaintiff, in confidence, and then publicly

329. *See* Ultramares Corp. v. Touche, 174 N.E. 441, 444 (N.Y. 1931).

330. Moore Charitable Found. v. PJT Partners, Inc., 217 N.E.3d 8, 14-19 (N.Y. 2023).

331. *Id.* at 22-23 (Singas, J., dissenting).

332. *Id*. at 15-16 (majority opinion).

333. *See, e.g.,* Olli Järviniemi & Evan Hubinger, *Uncovering Deceptive Tendencies in Language Models: A Simulated Company AI Assistant*, ARXIV 1-5 (Apr. 25, 2024), https://arxiv.org/pdf/2405.01576 [https://perma.cc/C5F3-3YGQ] (demonstrating that state-of-the-art language models engage in strategic deception in certain simulations without external pressure or instruction).

releasing it.[334] Should the courts hold the developer liable under a duty to take care against releasing such a model (i.e., a model that commits something like the "tort" of publicly disclosing private facts)?[335] Doing so could impose a significant economic burden on foundation-model developers, for obvious reasons. And it would require some aggressive judicial doctrinal development, since the underlying tort requires recklessness, at present, not mere negligence.[336] But the case for imposing such a duty, despite these concerns, is equally compelling.

In fact, in determining whether and how to enrich the developer's duty of care in negligence, there may be an argument for looking even beyond the criminal law and intentional torts, to bodies of law such as contract. Suppose that an AI developer releases a model that is deployed by a small business as a customer-service representative. The model malfunctions and makes an offer that some customer accepts, and on which she later justifiably relies. If the customer suffers economic loss in consequence, may she recover against the developer, if the developer failed to take reasonable care against such a malfunction? There is a plausible case for that proposition, especially since neither the law of contract nor the law of agency seems to hold the business liable for the model's malfunctioning.[337] But to effect such a recovery, tort law will need to recognize something like a duty to take care against causing economic loss by releasing a model that engages in a *contractual* malfunction (i.e., a model that malfunctions by "making" or "accepting" a contract offer). At present, an AI developer has no such duty in either tort or contract. Perhaps that should change.

Similarly, suppose a developer releases a model that malfunctions by repeatedly accessing some plaintiff's online database; the model does not cause tangible harm or data loss, but forces the plaintiff to expend considerable financial resources in order to understand the threat and protect itself. If the model were treated as an agent capable of "violating" the law, its action could conceivably be held to violate the Computer Fraud and Abuse Act (CFAA).[338] Tort law, by contrast, does not prohibit (or offer redress for) such actions.[339] Thus a court that wishes to provide

334. *Cf.* Nizan Geslevich Packin, *Generative AI Under Attack: Flowbreaking Exploits Trigger Data Leaks*, FORBES (Nov. 26, 2024), https://www.forbes.com/sites/nizangpackin/2024/11/26/generative-ai-under-attack-flowbreaking-exploits-trigger-data-leaks [https://perma.cc/P6KY-UWA3] (discussing "flowbreaking," a tactic that can cause models to leak confidential information without authorization).

335. Hill v. Nat'l Collegiate Athletic Ass'n, 865 P.2d 633, 644 (Cal. 1994).

336. *Id*.

337. Existing discussions of these issues tend to focus on the liability of the deployers, rather than developers, of algorithms. *See, e.g.*, SAMIR CHOPRA & LAURENCE F. WHITE, A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS 29-70 (2011); Matthew Oliver, *Contracting by Artificial Intelligence: Open Offers, Unilateral Mistakes, and Why Algorithms Are Not Agents*, 2 AUSTL. NAT'L U. J.L. & TECH. 45, 49-50 (2021).

338. 18 U.S.C. § 1030 (2024).

339. *See* Intel Corp. v. Hamidi, 71 P.3d 296, 302-09 (Cal. 2003).

redress against a developer who has negligently released such a model will need to borrow from the substance of the CFAA in order to articulate a novel tort duty of care on the developer.

How best to characterize the content of such duties of care, and whether their recognition is ultimately warranted, are difficult questions. The general point for now is that, because foundation models are polymathic and protean, they can injure a vast array of interests that are currently protected by the intentional torts, criminal law, contract law, and so on, but *not* by the law of negligence (or products liability, for that matter). For this reason, the developer's duty of care will need to borrow from the content of these other legal prohibitions if tort law is to afford adequate redress in a world filled with foundation models. Exploring how this might be done, by courts and legislatures, may become a pressing research program for torts scholars if the agentic capabilities of foundation models continue to improve.

B. Creating a Behavioral Malfunction Doctrine in the Law of Negligence

As discussed above, it is plausible that, if a model behaves in egregiously undesirable ways—for example, validating an evidently psychotic user's clearly false beliefs or actively encouraging a user to commit suicide—the plaintiff ought to be able to recover from the developer without establishing fault.[340] One way to grant such recovery would be to hold the developer vicariously liable for its model's "torts," but (as discussed earlier) there are doctrinal and normative difficulties with such a view,[341] and it would represent a very aggressive development of common-law doctrine that many courts are likely to be reasonably reticent to conduct.

There is a more incremental doctrinal development, within the law of negligence, that can do much of the same work: a plaintiff who establishes that an AI model has behaved in a clearly flawed way—that the model has, in an intuitive sense, *malfunctioned*—should be deemed to satisfy her burden of production on breach, such that she reaches the jury and the jury can permissibly find for her even if she does not adduce expert evidence or any other further evidence on breach.

There is precedent for such a *malfunction doctrine* within products liability: many states, in line with the *Restatement (Third)* of products liability, allow clear product malfunctions to serve as circumstantial evidence sufficient to meet the plaintiff's burden of production with respect to defect, such that the plaintiff may reach the jury.[342] This doctrine is itself understood as an analog, within products liability, of res ipsa

340. See *supra* Section II.A.2.
341. See *supra* Section II.C.
342. *See* GEISTFELD, *supra* note 35, at 91-103.

loquitur,[343] the doctrine of circumstantial evidence in the law of negligence that allows a plaintiff to satisfy her burden of production by showing that an instrumentality that was under the "exclusive control" of the defendant has caused her injury[344] or that simply inquires, more flexibly, into whether the accident was of a type "that ordinarily happens as a result of the negligence of a class of actors" of which the defendant was a part.[345]

The first formulation of res ipsa doctrine is not well placed to do much work in the context of foundation-model development, since ideas such as control do comparatively little to distinguish fact patterns that indicate developer negligence from those that do not. The second formulation does better. Nevertheless, trial courts and litigants could benefit from greater doctrinal specificity about the conditions under which model behavior is properly regarded as satisfying this description.

The intuitive idea of a behavioral malfunction goes some distance toward providing greater specificity. Leaving the operative idea of a malfunction without further specification might be defensible, at least to begin. Ultimately, however, further doctrinal specification might be warranted and desirable.

One such specification would draw on the idea of law-following AI[346]: when an AI model behaves in a manner closely comparable to a human agent, and in such a way that would be seriously unlawful and injurious (e.g., a crime or an intentional tort) if committed by a human, the "unlawfulness" of the agent's behavior might be deemed a malfunction so as to satisfy the plaintiff's burden of production on breach. Alternatively, or in addition, courts could hold that when a model behaves in a manner that violates a *clear* or *concrete* safety expectation held by its user or third parties, the model malfunctions in a way that satisfies the plaintiff's burden.[347]

Either specification of the malfunction doctrine will properly handle the sorts of cases that warrant its application. Both specifications would, for example, capture the intuitively plausible position that a model that causes a user to commit suicide or homicide by encouraging that user to do so, or by validating that user's clearly delusional beliefs about the murderous intentions of his intimates, has egregiously malfunctioned in a manner that should suffice to relieve the plaintiff of the burden to adduce further evidence of breach.

---

343. See *supra* note 192 and accompanying text.

344. RESTATEMENT (SECOND) OF TORTS § 328D (AM. L. INST. 1965).

345. RESTATEMENT (THIRD) OF TORTS: LIAB. FOR PHYSICAL & EMOTIONAL HARM § 17 (AM. L. INST. 2010).

346. *See generally* O'Keefe et al., *supra* note 323 (championing and exploring this idea).

347. *Cf.* Clayton J. Masterman & W. Kip Viscusi, *The Specific Consumer Expectations Test for Product Defects*, 95 IND. L.J. 183, 183 (2020) (proposing a "specific consumer expectations test [that] would apply to any product or product component for which consumers have clear, articulable ex ante expectations about the function of the product").

Under any such formulation, the malfunction doctrine should also help to address the enormous asymmetry in technical expertise and knowledge between frontier AI developers and basically all other interested parties, including regulators, courts, private persons who might become injured plaintiffs, and even academic experts in machine learning.[348] Because of this asymmetry, injured plaintiffs may find it difficult to locate expert witnesses with the competence and willingness to testify on their behalf. The difficulty is compounded by the dense social, professional, and (in some cases) financial connections between industry and technical experts outside of it. Many of the leading independent organizations that red team foundation models for dangerous capabilities were founded, and are largely staffed, by alumni of the frontier labs.[349] It is reasonable to suppose that many of these technical experts retain substantial financial exposure (in the form of equity and options) to the firms at which they worked, along with friendships and other social connections.

Moreover, since these independent organizations currently depend on the voluntary cooperation of the frontier labs for access to their AI systems, they have powerful incentives to refrain from any behavior that would jeopardize this access. Academics have different kinds of incentives to remain in industry's good graces, but in their case too the incentives are powerful.[350] In consequence, even if plaintiffs' attorneys have plentiful access to the evidence necessary to develop their cases, they might sometimes face serious difficulties in accessing the expertise necessary to contextualize this evidence for judges and juries. While there is reason to hope that plaintiffs will surmount these difficulties (thereby creating societal benefits involving information production and the like), a malfunction doctrine that relieves the ordinary burden of production on breach in a sufficiently predictable fashion should help to induce plaintiffs to bring meritorious claims even when they cannot overcome these challenges or do not wish to invest in doing so.

348. *See, e.g.*, Nestor Maslej et al., *Artificial Intelligence Index Report 2024*, STAN. INST. FOR HUM.-CENTERED AI 58 (2024), https://hai.stanford.edu/assets/files/hai_ai-index-report-2024-smaller2.pdf [https://perma.cc/3WKJ-4L54] (noting that nearly three-quarters of foundation models originated from industry in 2023); *id.* at 328 (providing statistics on the intensification of brain drain from academia to industry).

349. To give just one example, several employees of METR, perhaps the leading independent red-teaming nonprofit, previously worked at frontier AI companies such as Google DeepMind and OpenAI. *See About METR*, METR, https://metr.org/about [https://perma.cc/58TL-SGNE].

350. As Arvind Narayanan, a machine learning professor at Princeton, has put it, "[t]he agenda-setting power of tech companies means that CS academia is—and always has been—captured by the industry. You can turn down industry funding but you can't chart a truly independent research direction, because the success of research depends on community acceptance." Arvind Narayanan (@random_walker), X (Jan. 23, 2025, 7:27 AM), https://x.com/random_walker/status/1882404836601450777 [https://perma.cc/M65T-AP8S].

Conclusion: The Limits of the Common Law, the Externalities of Tort Liability, and the Necessity of *Ex Ante* Regulation

The preceding Part argued that certain limitations of existing negligence doctrine can be addressed through incremental judicial development of the law. With such development, the law of negligence can deftly and flexibly handle many of the challenges that frontier AI systems will pose for the tort system. The case for replacing or supplementing ordinary negligence liability with salient doctrinal alternatives, such as strict liability for abnormally dangerous activities and vicarious liability, is much weaker than many scholars writing on AI liability have supposed. The case for supplementing it with products liability, a complex mix of negligence-like and stricter liability rules, is stronger. But even that intervention faces underappreciated difficulties, and in any event products liability cannot cover certain important pathways through which foundation-model development might cause serious harm, such as internal deployments that result in models escaping from their testing environments and causing external mischief. The law of negligence elegantly and naturally handles the full range of relevant risks.

But we have also encountered certain limitations of the common-law tort system—and, indeed, *externalities* of it—that transcend the choice between common-law liability rules. No amount of incremental doctrinal development, of the sort that courts are competent to perform, will adequately address these externalities.

First, the prospect of tort liability can disincentivize developers from investigating and disclosing (to the public, government, and suitable third parties) the harms that their activities may have caused.[351] This perverse-incentive effect is, in one respect, worse under strict liability than negligence; in another respect, it is worse under negligence than strict liability. But *any* form of compensatory liability will tend to have an effect of this kind.

This is a perfectly general phenomenon that has nothing in particular to do with AI development. Nevertheless, it is distinctively *troubling* in the case of AI development. Because of the extraordinarily dense concentration of leading technical expertise and cutting-edge information and technology within the companies developing frontier AI, our society is relying on those companies themselves to understand the risks of frontier AI (and how these risks might be mitigated) to an extraordinary degree. And some of the most severe risks that frontier AI might pose—such as risks arising from the prospect that misaligned AI systems might evade and subvert human control, or risks involving the facilitation of biological weapons creation—are not yet well understood. Thus, it is plausible that acquiring better information and evidence about the risks and harms of

351. *See supra* Section II.A.5.

frontier AI is the most urgent and important task of AI governance at the technological frontier.[352] Thus, it is quite disturbing that the common law of torts—the only existing legal institution that governs AI development in any general way—can be expected, in certain important ways, to hinder this task.

Second, the prospect of tort liability cannot adequately incentivize care with respect to harms so enormous that compensating for them would render a developer insolvent.[353] For similar reasons, tort liability might actively *weaken* a developer's incentives for care. Suppose that a developer, *D*, causes a harm, *H*, so enormous that *D* will be rendered insolvent if forced to compensate for it. In the interim between learning of *H* and litigating the resulting tort case to judgment, *D* may be rationally incentivized to gamble for resurrection, so to speak, by engaging in further risky activities that could make it enough money to remain in business even after paying compensation for *H*.[354] After all, the more likely it is that *D* will be required to compensate for *H,* the lower will be the expected disvalue to *D* of risking causing *another* harm. Perhaps this perverse incentive can be mitigated by a rule of negligence rather than strict liability. (Under a negligence rule, *D* is less likely to be required to compensate for *H*; moreover, *D* might reasonably expect that behaving riskily going forward will prejudice the jury adjudicating whether it was negligent in causing *H*.) But *any* exposure to liability will tend to generate this effect. Frontier AI developers are already racing to create generally intelligent, vastly superhuman AI. If such a developer should fall into financial distress, racing even faster might prove a rationally self-interested—but socially catastrophic—gamble for resurrection.[355]

It is possible that legislatures can address the preceding limitations and externalities of the tort system, at least to some extent, by modifying

352. *See* Dean W. Ball & Ketan Ramakrishnan, *Entity-Based Regulation in Frontier AI Governance*, CARNEGIE ENDOWMENT FOR INT'L PEACE (July 7, 2025), https://carnegieendowment.org/research/2025/07/artificial-intelligence-regulation-united-states [https://perma.cc/2W9L-UG2F] ("How to identify and measure [the] risks [of frontier AI], when exactly they will arise, and how to respond to them are all matters of considerable uncertainty and disagreement. One of the chief tasks of a frontier AI regulatory regime—arguably the chief task, at least for now—is to put society in a position to reduce such uncertainty and disagreement by enabling policymakers and the public to better understand the nature and magnitude of the risks posed by cutting-edge AI development.").

353. This issue is likely less severe under negligence liability than under strict liability. But it is still severe, given litigation costs, legal uncertainty, and the practical reality of adjudicative error. *See supra* Section II.A.3.

354. On this phenomenon, see generally Michael C. Jensen & William H. Meckling, *Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure*, 3 J. FIN. ECON. 305, 334-43 (1976).

355. The point holds, in qualified form, even if a developer can expect to receive a bailout from government institutions, since the developer's officers and directors (who are, in the final analysis, the agents actually making the relevant risky decisions) can generally expect to suffer much greater financial loss (and perhaps greater loss of other kinds, such as loss of social status, reputational capital, and political power) if the firm is rendered insolvent.

the common law in ways that lie beyond the competence and legitimacy of the courts. Most importantly, legislatures might consider:

- *Penalties and supracompensatory damages for failure to investigate or disclose tortious harms.* Legislatures could lay down that, if a developer is successfully sued for tortiously injuring a plaintiff, and the developer did not discover and disclose (to the public or government) the harmful incident—or, alternatively, did not take reasonable measures to discover and disclose the incident—the developer is liable not only to compensate the plaintiff but also to pay the plaintiff or the government additional, supracompensatory damages.
- *Penalties and supracompensatory damages for failure to employ independent investigators, auditors, and evaluators.* Legislatures could lay down that, if a developer is successfully sued for tortiously injuring a plaintiff, and the developer did not employ competent, credibly independent third parties to periodically investigate risks and harms of the relevant kind, the developer is liable not only to compensate the plaintiff but also to pay the plaintiff or the government additional, supracompensatory damages.
- *Officer and director liability for failure to employ third-party investigators, auditors, and evaluators in the aftermath of catastrophe.* [356] A legislature could lay down that, if a developer tortiously causes certain catastrophic harms, and the developer did not employ competent, credibly independent third parties to periodically investigate risks and harms of the relevant kind, then the developer's officers and directors are personally liable to compensate for the harms in question.
- *Penalties, supracompensatory damages, or officer and director liability for failure to utilize a safety precaution utilized by another comparable AI developer.* A legislature could lay down that, if a developer tortiously causes certain catastrophic harms, the developer did not employ a concrete safety precaution utilized by at least one other comparably situated developer, and the failure to utilize that precaution is plausibly a but-for cause of the catastrophic harms in question, then the developer is liable to penalties or supracompensatory damages and the developer's officers and directors are

356. For a similar proposal in a different context, *see generally* Kyle D. Logue, W. Robert Thomas & Jeffery Y. Zhang, *Sanctioning Negligent Bankers*, 78 STAN. L. REV. 607 (2026).

personally liable for compensatory damages, penalties, and/or supracompensatory damages.

The desirability and optimal design of such legislative possibilities are certainly worth consideration. In the final analysis, however, it seems unlikely that any such modification of the existing liability regime can adequately counteract the perverse incentives that exposure to liability, by its nature, creates for AI developers. That is especially so at the technological frontier, where it is critical to investigate the distinctive risks of AI development to mitigate the severity of potential harms so large that they threaten developer insolvency.

For this reason, among others, an optimal system of frontier-AI governance will need to make considerable use of *ex ante* regulatory requirements and institutions alongside *ex post* mechanisms such as liability and penalties.[357] As frontier AI models continue to grow more powerful, and highly dangerous capabilities emerge, it will become increasingly untenable to maintain that such *ex post* mechanisms can serve as the principal institutional strategy for addressing their dangers.[358] Ultimately, frontier AI developers must be legally required to investigate and mitigate the risks that their activities pose or may pose. And independent parties, whether private firms or government officials, must be legally empowered to ensure that AI developers are performing this task with sufficient diligence and competence, before those risks ripen into harm.

357. For a brief overview of *ex ante* regulatory possibilities, see Ball & Ramakrishnan, *supra* note 352 ("At a minimum, [frontier AI] regulation might involve transparency requirements: obligations on covered developers to disclose (to the public, the government, and/or other appropriate parties) certain salient features of their riskiest activities (such as training, safety testing, and assessment and monitoring of insider threats and high-stakes internal deployments), and to disclose when certain particularly risky capabilities or propensities have plausibly emerged in their models or systems . . . . At the other end of the spectrum, regulation might require the sort of highly intensive government engagement that characterizes nuclear safety regulation.").

358. For a broadly similar view, see generally Daniel Schwarcz & Josephine Wolff, *The Limits of Regulating AI Safety Through Liability and Insurance: Lessons from Cybersecurity,* 95 FORDHAM L. REV. (forthcoming 2026). For a different view, see Weil, *supra* note 170, *passim*. See also Ketan Ramakrishnan, *Tort Law Is the Best Way to Regulate AI*, WALL ST. J. (Sep. 24, 2024), https://www.wsj.com/opinion/tort-law-is-the-best-way-to-regulate-ai-california-legal-liability-065e1220 [https://perma.cc/4KMM-4KXK], where I suggested that "[s]ince our understanding of AI safety is nascent, and highly dangerous capabilities haven't yet emerged, instituting command-and-control regulation may be premature," and that "we should regulate AI development by building on the common law of torts, one of our oldest legal institutions," instead. Unfortunately, while our understanding of AI safety remains nascent, highly dangerous capabilities are now emerging.