OVERCOMING UNFORESEEABLE AI ERRORS: A PROPOSAL

Mitchell Chervu Johnston[*] *& Brian Matejek*[†]

Artificial Intelligence does not simply make errors; it makes unusual errors. Some AI errors are thus seemingly truly unforeseeable. And because traditional tort law categories embed foreseeability within them, these unforeseeable errors vex attempts to classify AI within traditional doctrinal boxes. We argue that it is this "foreseeability gap" that generates many of the hard problems of applying tort law to AI systems. While the traditional response to unforeseeability (and one that appears in previous literature on AI) is insurance, we argue that the foreseeability gap can be better closed by careful attention to the way in which AI systems work. In particular, AI systems are fundamentally prediction algorithms, and, in some cases, they fail for known reasons, even if the resulting errors are themselves unusual. This Article provides a taxonomy of these failures and proposes a novel approach to liability rooted in the reason *for the AI failure. This proposal, we argue, balances fairness to developers and users of AI with the goal of compensating victims of accidents. Moreover, it is flexible enough to adjust to advances in AI technology and AI research. The result is a proposal that overcomes various shortfalls with prior approaches.*

[*] Assistant Professor of Law, Boston College Law School.

[†] Computer Science Laboratory, SRI International. Thanks to Dhruv Aggarwal, Zach Fields, Nathaniel Donohue, James Grimmelmann, Patrick Hulin, Peter Kallis, and Shelly Simana for helpful conversations and comments on the draft. Opinions expressed are the authors' own.

INTRODUCTION

AI companies and developers promise to bring a revolution to how we work and live. Even if we discount the more utopian of these claims, it is clear that AI systems are already in wide use for both work and recreation. And with the wide deployment of a novel technology, questions then turn to how the law should police the inevitable accidents that will occur.

Legal scholars and policy makers have already provided a number of answers to these questions. One approach is *ex ante* regulation of AI uses, an approach that is already being deployed by European regulations via the AI Act.[1] But even if *ex ante* regulation is deployed (and even if it is wise to regulate emerging technology in this way), accidents will still occur. Scholars have thus explored the question of what scheme of *ex post* liability makes sense. Arguments have been made for and against traditional negligence, strict liability, products liability, animal liability, vicarious liability, enterprise liability, or even to treat AI models as autonomous individuals.[2]

---

[1] *See* Lilian Edwards, *The EU AI Act: A Summary of its Significance and Scope*, ADA LOVELACE INST. (Apr. 2022), https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf

[2] *See, e.g.*, RYAN ABBOTT, THE REASONABLE ROBOT: ARTIFICIAL INTELLIGENCE AND THE LAW (2020) (arguing that the law should pursue parity between humans and artificial agents including in tort law); Miriam C. Buiten, *Product Liability for Defective AI*, 57 EUR. J.L. & ECON. 239, 240-41 (2024) (arguing that product liability law must adapt its notion of a defect to account for "how AI systems disrupt the traditional balance of control and risk awareness between users and producers"); Bryan H. Choi, *AI Malpractice*, 73 DEPAUL L. REV. 301, 306 (2024) (suggesting a professional malpractice standard for AI modelers but noting that "for certain aspects of AI work, there is greater occasion to consider alternate regimes such as strict liability or ordinary negligence"); Mihailis E. Diamantis, *Employed Algorithms: A Labor Model of Corporate Liability for AI*, 72 DUKE L.J. 797, 805 (2023) [hereinafter Diamantis, *Employed Algorithms*] ("[T]he Labor Model maintains that if a corporation and an algorithm share the hallmarks of an employment relationship—substantial benefit and substantial control—then the algorithm should be deemed an employed algorithm for whose harms the corporation could be liable."); Mihailis E. Diamantis, *Reasonable AI: A Negligence Standard*, 78 VAND. L. REV. 573, 580-81 (2025) [hereinafter Diamantis, *Reasonable AI*] (proposing a hybrid negligence standard); F. Patrick Hubbard, *"Sophisticated Robots": Balancing Liability, Regulation, and Innovation*, 66 FLA. L. REV. 1803, 1862-65 (2014) (considering analogies to respondeat superior, children, animals, and abnormally dangerous activities); Anat Lior, *AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy*, 46 MITCHELL HAMLINE L. REV. 1043, 1046 (2020) (considering analogies to "products, animals (domesticated and wild), slaves, electronic persons, and a grander analogy to the aviation and vaccination industries"); Chinmayi Sharma, *AI's Hippocratic Oath*, 102 WASH. U.L. REV. 1101 (2025) (proposing professionalizing AI engineering and asking engineers to comply with mandatory and technical guidelines); Steven Shavell, *On The Redesign of Accident Liability for the World of Autonomous Vehicles*, 49 J. LEGAL STUD. 243, 244-45 (2020) (arguing that traditional strict liability and fault-based liability rules do not perform well to regulate accident risk due to autonomous vehicles and suggesting an alternative rule of strict liability with damages paid to the state); David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117, 149 (2014) (suggesting common enterprise liability); Christiane

While many of these approaches are compelling, they stem from a common instinct to fit artificial intelligence into an existing doctrinal framework. That instinct is ordinarily healthy: existing frameworks like negligence have long pedigrees in part because of their adaptability and flexibility.[3] Moreover, stated at a high level of generality, the various traditional doctrinal tools arguably exhaust the menu of options.[4]

Nevertheless, in this Article we depart (albeit modestly) from this instinct.[5] Traditional frameworks of tort liability implicitly embed various social understandings about the types of errors that others make.[6] They assume not only that humans err but that these errors follow a regular distribution that other actors can understand. Normal law firm associates do not typically invent citations out of whole cloth, AI "associates" do.[7] By now, of course, lawyers

---

Wendehorst, *Strict Liability for AI and other Emerging Technologies*, 11 J. EUR. TORT L. 150, 179-80 (2020) (proposing a system of strict no-fault liability for physical harms caused by AI that applies strict liability based on the nature of the physical risk); Hilyard Nichols, Note, *The First Byte Rule: A Proposal for Liability of Artificial Intelligences*, 15 W&M BUS. L. REV. 189, 194 (2023) (suggesting an approach analogous to animal liability)

[3] *See* Hubbard, *supra* note 2, at 1806 (critiquing proposals for fundamental changes to the legal system in response to advances in robotics on the grounds that the current system is likely to allocate risks fairly).

[4] That is, liability can be based on fault or not, fault can be flexibly defined, and the law can attribute fault vicariously or otherwise. Moreover, the law offers a menu of flexibly tailored standards of care for fault such as professional standards or ordinary care standards. Once the high level design choice is made, the real trouble is figuring out how to decide what faults count and whose faults they are.

[5] *See* Rebecca Crootof, *Autonomous Weapon Systems and the Limits of Analogy*, 9 HARV. NAT'L SEC. J. 51, 55-59 (2018) (critique the tendency to analogize autonomous weapons systems to existing legal categories on the grounds that the relevant analogies erase legally relevant aspects of such systems); *see also* Ignacio N. Cofone, *Servers and Waiters: What Matters in the Law of A.I.*, 21 STAN. TECH. L. REV. 167, 169-71 (2018) (developing a three factor framework that would sort AI into the appropriate legal analogy).

[6] *See infra* Section II.A.

[7] Several lawyers (and judges) have been discovered filing court documents with "hallucinated" citations—citations to fictional cases. *See* Zach Warren, *GenAI Hallucinations Are Still Pervasive In Legal Filings, But Better Lawyering is The Cure*, THOMPSON REUTERS (Aug. 18, 2025), https://www.thomsonreuters.com/en-us/posts/technology/genai-hallucinations/; Daniel Wu, *Federal Judges Using AI Filed Court Orders With False Quotes, Fake Names*, WASH. POST (Oct. 29, 2025), https://www.washingtonpost.com/nation/2025/10/29/federal-judges-ai-court-orders/. In some cases, it seems these issues have arisen from a blind trust in the algorithm to write portions of documents. Though these errors have an amusing quality to them, it is not the case that lawyers have never delegated the task of researching and writing motions to other intelligences; rather it is typical for junior lawyers to draft documents and ensure the citations are correct. It is, of course, still the case that it is the duty of the lawyer signing the papers to certify that the material contained within, but outside of extreme cases, it seems fair to say that human associates do not *invent* cases wholesale. Senior lawyers can thus trust junior associates to cite real cases (even if they may do so inaccurately). But when an algorithm that seems to write and research like a person stands in for a human employee it changes the management challenge itself. In this case the challenge is not so difficult to overcome (lawyers should learn to cite check AI-written

should be on notice about the problem of hallucinations and, as a result, we think it is fair to treat them as expected and foreseeable by reasonable practitioners. But as AI usage has spread new issues have continued to emerge. New users using new AI systems in new ways promise to create unseen errors that, as AI systems are empowered to act in the real world, will likely create unforeseen harms.[8]

Proposals that attempt to fit artificial intelligence into traditional categories, we think, overlook that artificial intelligence regulation is tricky, in part, because the structure of AI errors is *unusual*.[9] AI systems are said by some legal commentators to be a "black box" (though we think "opaque" or "of uncertain variability" are superior descriptions of the problem), true, but so are the minds of others.[10] Unpredictability, incomprehensibility, and oddity are the salient features of AI harms that call out for a response.[11]

---

briefs closely) but the discontinuity between what to expect from fallible humans and what to expect from fallible machines still generates friction as the new technology is adopted.

[8] *See, e.g.*, Rohin Shah et al., *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals*, ARXIV (Nov. 2, 2022), https://arxiv.org/abs/2210.01790 ("*Goal misgeneralization* refers to this pathological behaviour, in which a learned model behaves.").

as though it is optimizing an unintended goal, despite receiving correct feedback during training

[9]Bruce Schneier & Nathan E. Sanders, *AI Mistakes Are Very Different From Human Mistakes*, IEEE SPECTRUM (Jan. 13, 2025), https://spectrum.ieee.org/ai-mistakes-schneier ("It seems ridiculous when chatbots tell you to eat rocks or add glue to pizza. But it's not the frequency or severity of AI systems' mistakes that differentiates them from human mistakes. It's their weirdness. AI systems do not make mistakes in the same ways that humans do."); *cf.* Diamantis, *Reasonable AI*, *supra* note 2, at 598-600 (noting that some AI errors are "dumb" from the perspective of humans, collecting examples, and concluding that the ordinary person is a poor benchmark for AI reasonableness). In an interesting forthcoming Note Trent Kannegieter suggests a similar approach to ours. *See* Trent S. Kannegieter, *Nondeterministic Torts: A Mechanistic Approach to Large Language Model Tort Liability*, 135 YALE L.J. (forthcoming 2026). Like us, Kannegieter is interested in mechanisms and like us he identifies the important problem as the ways in which the structure of LLMs cause them to behave in ways that do not neatly map onto traditional problems. Kannegieter, however, roots his analysis in *nondeterminism* while we are focused on the *unusual* nature of non-deterministic outputs and the implications for foreseeability.

[10] In particular, the "black box" framing suggests that we cannot see inside the mechanism. But the puzzle of AI isn't really that sort of problem. Even if the weights of the model are totally known, the output may be hard to explain. We can thus know the input and have complete knowledge of the numerical transformations that will occur and still produce outputs of variable explainability. Indeed, the thesis of this Article is that precisely because we can understand part of what is happening under the hood of such models, we can use this knowledge to reach better decisions about liability.

[11] Previous scholarship has often characterized the issue of AI errors as a question of predictability, explainability, and diagnosis. *See, e.g.*, Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 889, 891 (2018) (arguing that machine learning algorithms and their black box nature frustrate efforts to assess intent and establish causation); Choi, *supra* note 2, at 304 ("Yet, with modern AI techniques such as deep neural networks, there is broad consensus that failures are difficult to diagnose or explain in human-understandable terms. Thus it can be difficult to discern when law should exact remedies

Consider three examples of AI errors. First, researchers at OpenAI discovered that an AI trained to play a boat racing video game discovered a degenerate solution in which it would drive the boat in a circle and "[d]espite repeatedly catching on fire, crashing into other boats, and going the wrong way on the track" would achieve a high score in the game.[12] Second, in a real world situation, a Tesla operator "Summon"-ed the car which proceeded to drive onto a tarmac in its self-driving mode and crash into a parked jet.[13] Third, consider the phenomenon of adversarial examples. These are cases in which a slight distortion to an input that renders it indistinguishable to a human can fool an AI classification system.[14] In the former two examples, the AI system attempting to complete its task errs in a way that its developers did not intend and that a human asked to accomplish the same task almost certainly would not. In the latter example, a human attacker is able to trick the AI system via a method that would not work on a human counterparty and that seems highly unusual.

As a result, proposals to treat AI as a product, for example, and therefore to apply product liability doctrines to it must grapple with the way in which products liability doctrines allocate liability between manufacturers and users based on understandings about the ways in which products are predictably and unpredictably used and misused in ways that harm others.[15] Similarly, doctrines

---

from AI modelers versus when law should let costs lie where they fall."); Anat Lior, *Holding AI Accountable: Addressing AI-Related Harms Through Existing Tort Doctrines*, U. CHI. L. REV. ONLINE *1, *3-*5 (Nov. 2024) ("The primary challenge which stands at the center of the AI liability debates is the 'black box' issue. This refers to the fact that an AI-based algorithm's decision-making progress is opaque and unknown to the user or programmer of the algorithm. This hampers a core notion standing at the heart of tort law—foreseeability and predictability.").

These authors and others observe that because AI is unpredictable and its errors it is both difficult to deter them (their creators and users cannot predict them) or to attribute causation to this or that action (the explainability/diagnosis issue). To be sure, the unpredictability that Bathee, Choi, and Lior (and others) emphasize surely sweeps in issues of oddity. *See* Diamantis, *Reasonable AI*, *supra* note 2, at 594-95 (observing that AI's ability to find novel solutions to problems relies on its ability to engage in unforeseen behavior). The difference is one of emphasis and where it leads us. Our focus on the oddity of these errors pushes us to be interested in whether odd errors can sometimes be traced to a set of mechanisms that can become known over time thus permitting a level of foreseeability.

[12] Jack Clark & Dario Amodei, *Faulty Reward Functions in the Wild*, OPENAI (Dec. 21, 2016), https://openai.com/index/faulty-reward-functions/.

[13] Justin Westbrook, *Tesla Model Y Operator Appears to "Summon" Car Straight Into a Parked Jet*, MOTORTREND (Apr. 22, 2022), https://www.motortrend.com/news/tesla-model-y-summon-aircraft-crash-cirrus-vision-jet.

[14] Ian Goodfellow et al., *Attacking Machine Learning with Adversarial Examples*, OPENAI (Feb. 24, 2017), https://openai.com/index/attacking-machine-learning-with-adversarial-examples/ (proving several examples). The authors noted at the time of writing it was very difficult to design a system to defend against adversarial examples and the two techniques available still could be broken by an attacker. *Id.*

[15] Of the existing literature, work by Mariam Buiten, Alexandre de Streel, and Martin Peitz, comes closest to our approach. Mariam Buiten, Alexandre de Streel & Martin Peitz, *The Law and*

of vicarious liability for torts committed by employees embed assumptions about the ways in which human beings err and thus when it is fair to attribute their errors to others on the grounds that someone else should (or could best) have controlled those errors.[16] Digital employees may fail to satisfy these assumptions, undercutting their fairness rationale.[17]

In practice principles of responsibility undergirded by foreseeability (as expanded by courts over time) suffice to protect innocents from injuries caused by the errors of others. But it does not follow that this will be so in the face of unpredictable new technologies. Indeed, as we argue here, what does so (in practice) is what we term the *human-fallibility assumptions* of tort law—that the boundaries of tort law are drawn with both the limitations and frailties of human beings *and* the social knowledge on behalf of *others* of what such limitations are.[18] Taking these assumptions as a starting point, we can then see one view of why tort law treats other objects of regulation, from products to wild animals, the way it does.

The problem of harms caused by artificial intelligence, we argue, is generated by what we term the *foreseeability gap*.[19] To the extent that accidents generated by artificial intelligence seem random and unpredictable and thus unlike traditional accidents, traditional principles are in a tough spot. In truly unexpected cases, losses lie where they will.[20] Such cases, however, are assumed to be rare and unusual (and, in many cases, insured). The puzzle of AI, as David Vladeck observed in an early article on the matter (albeit one focused on an

---

Economics of AI Liability, 48 COMP. L. & SEC. REV. 1, 5-7 (2023) (acknowledging that the complexity, autonomy, and unpredictability of AI systems confound traditional fault-based liability). We thus do not wish to claim falsely that previous work has ignored this problem. Instead, we are arguing that the unpredictability of AI errors is critically important to satisfactory adaptation of tort law (and regulatory law in general) to AI accidents. Moreover, we aim here to provide a more detailed solution to this problem based, in part, on a deeper technological analysis of these systems.

[16] *See infra* Section II.A.

[17] *See infra* Section I.B.

[18] *See infra* Section II.A.

[19] Mihailis Diamantis has advanced a related notion he terms the "AI Accountability Gap." Diamantis, *Reasonable AI*, *supra* note 2, at 581-86. This refers to the difficulty he observes in applying traditional doctrines of responsibility to locate a proper defendant. *Id.*

[20] For tort law, this is simply a claim that would fail at some point on either duty, breach, or proximate cause, all of which include some consideration of foreseeability. *See infra* Section II.A. The law of contracts also contains a trace of the same notion insofar as it releases a party from performance in the case of impracticability, unforeseeable destruction of a thing necessary for performance, or frustration reflects a similar principle insofar as disgrace is conditioned on the occurrence of an "event the non-occurrence of which was a basic assumption on which the contract was made." *See* RESTATEMENT (SECOND) OF CONTRACTS §§ 261, 263, 265 (AM. L. INST. 1981). Here we see that where an unforeseeable event occurs losses will lie where they may. *Cf. id.* § 268 (noting that a claim of discharge based on these rules will also discharge prospective duties on the part of the counterparty).

earlier paradigm) is that traditional approaches based (implicitly and explicitly) on foreseeability struggle to cover this gap.[21] Confronted with this problem, one can, of course, deny the issue and claim that existing doctrines can indeed fill the gap. Alternatively, one can proceed down two orthogonal dimensions. First, it can be argued that the gap is not so broad as it appears and that indeed existing categories suffice to narrow it to an acceptable threshold. Second, one can discuss what fills the gap, if it exists, and choose between various schemes of strict liability or insurance (or, perhaps, no liability) to address the problem.

Our Article focuses on the first dimension, albeit with a twist. Taking inspiration from the insight about human-fallibility, we propose an approach that is focused on the *cause* of a particular error and on who and how that error is best controlled.[22] Potential developer liability is triggered by the ability to trace the error to a particular cause of failure. Then with this approach in mind, liability can be allocated between developers and users of AI systems via a system of warnings and notice. We thus believe that the foreseeability gap can be narrowed but with the wrinkle that foreseeability is evaluated at a somewhat higher level of generality.

---

[21] Vladeck, *supra* note 2, at 121. Vladeck's approach is influential precisely because it is careful to not dive too deeply into any one technology. Still, the analysis draws heavily upon the example of autonomous vehicles to motivate its conclusions though Vladeck clearly has broader systems in mind. *Id.* at 143-46 (discussing an autonomous vehicle example as a simpler case of the "HAL question" inspired by the AI system in *2001: A Space Odyssey* that starts to attack its human companions out of a belief that they are jeopardizing its mission).

[22] Our approach here is most similar to David Lehr and Paul Ohm's work urging legal scholars to understand the way in which machine learning works. David Lehr & Paul Ohm, *Playing With the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653 (2017). Lehr and Ohm offer a description of the process of building machine learning algorithms and convincingly argue that by understanding the process of building these models, legal scholars can sharpen their analysis of the harms and benefits of machine learning systems. *Id.* at 656-58. Our proposal thus builds upon this analysis by identifying what we see as the problem vexing analysis of AI systems in tort law and also by focusing on the *way* in which these built algorithms fail. We thus agree with their suggestion that more focus on the details of machine learning can sharpen our analysis of this space.

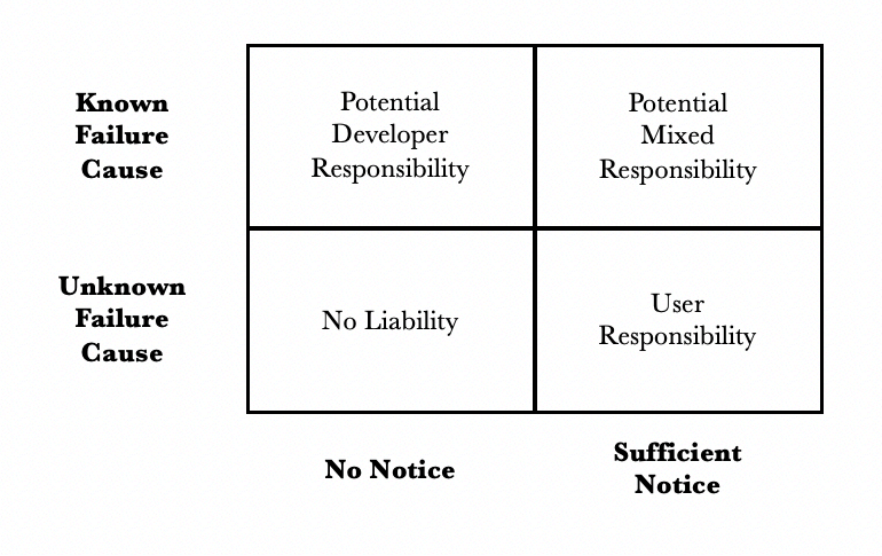

Figure 1: Liability Under the Model

Moreover, because one of us has a technical background working with artificial intelligence, we believe that we can offer a nuanced application of these traditional legal ideas to difficult technical problems. Indeed, the taxonomy of AI error causes below is, we believe, a contribution to the legal literature on its own.

Importantly, the taxonomy below did not emerge from a first principles analysis of AI systems. Instead, researchers observed the odd errors we discussed and over the years have endeavored to explain them (and develop defenses against them). As new errors emerge, the cycle repeats itself. A quick search on arXiv, the pre-print repository for computer science papers, shows 179 papers including the term "hallucinations" before 2020. The majority of these papers deal with convolutional neural networks trained to "hallucinate" new content into input images (e.g., to fill a hole or remove an occluding object). Since the start of 2021, there are 5,201 papers on hallucinations, most discussing the failures of large language models to ground themselves in facts.[23] Approaching the problem at this level allows the law to draw on the state of the art (and possibly encourage its development).

[23] As of January 22, 2026. The number of papers discussing hallucinations will inevitably grow until the problem becomes "solved" or a new shiny tool replaces large language models. Large language models use a transformer architecture. The groundbreaking transformer paper was published at NeurIPS in December 2017. *See* Ashish Vaswani et al., *Attention Is All You Need*, 31ST CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (2017). Chat-GPT was released to much fanfare in November 2022. The number of papers on arXiv discussing "hallucinations" (again, including other uses of the word) published before the start of 2023 was only 385.

To be sure, our proposal ultimately closely resembles traditional doctrines and, in many cases, reaches the same results. These resemblances are unsurprising. There are a number of fixed points in which AI regulation so closely resembles traditional tort problems that it's hard not to conclude that the cases should come out the same way. An autonomous vehicle (AV), for example, could be so carelessly designed that the manufacturer should be liable under traditional product liability principles.[24] Similarly, an employer who utilizes digital employees to screen resumes should be liable under traditional principles if those digital employees discriminate in obvious ways.[25] Indeed, we believe that any satisfactory proposal to regulate AI using tort law should be continuous in these "boundary cases."[26] Where AI acts like and falters akin to an object of traditional legal analysis we see no reason not to apply traditional legal principles. If an AI dog bites others reliably it seems sensible to apply the settled rules for dog bites.

Our proposal is thus intended to fill a gap rather than replace traditional doctrines when they apply neatly.[27] Indeed, we believe that proposals to regulate artificial intelligence systems should be "continuous at the boundaries" meaning that where AI behaves and err like a traditional category of legal regulation traditional rules should apply. The issue with AI governance does not arise in these cases but rather because AI can function in a way that resembles multiple potential categories and because the hard cases it creates flummox traditional approaches.

Our aim here is to make three contributions that will enable more general insights. First, we aim to provide the framework of an approach that can address both general-purpose AI models and specialized models. Though our errors-based approach will work out differently in different cases, we think a comprehensive solution to this problem may lie in an approach that encompasses

---

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

[25] Diamantis, *Employed Algorithms*, *supra* note 2, at 824-30 (arguing that a contrary result will create a liability loophole that allows corporations to defeat liability via automation that they would be liable for under traditional principles). Closely related are concerns about algorithmic price fixing that allow conspirators to avoid liability by hiding their collusion within the code of an algorithm. In such cases, commenters have suggested that if it is illegal for a human to do it is illegal for an algorithm to do. *See* Maureen K. Ohlhausen, *Should We Fear The Things That Go Beep In the Night? Some Initial Thoughts on the Intersection of Antitrust Law and Algorithmic Pricing*, FED. TRADE COMM'N 10-11 (May 23, 2017), https://www.ftc.gov/system/files/documents/public_statements/1220893/ohlhausen_-_concurrences_5-23-17.pdf ("Is it ok for a guy named Bob to collect confidential price strategy information from all the participants in a market, and then tell everybody how they should price? If it isn't ok for a guy named Bob to do it, then it probably isn't ok for an algorithm to do it either.").

[26] Diamantis, *Employed Algorithms*, *supra* note 2, at 843 (proposing that where traditional legal theories would permit recovery his proposal should not apply and instead plaintiffs should proceed under the traditional theory).

[27] *See infra* Part IV.

both types of cases (especially as specialized models are built off of generalized ones).[28] Second, we aim to do so in a way that draws upon technical insights into how these systems work and how they go awry in a way that moves down one level of generality from broad questions about fault to more specific questions of how to tackle certain concrete cases. Third, we aim here to develop a framework that can be robust to future developments. This final point is of particular resonance for our focus as the problem of "unexpected" AI errors may indeed become moot if AI development stalls and current models are deployed throughout society.[29] If so, formerly unexpected errors may become familiar and thus fit more neatly into traditional categories.[30] Hallucinations of sources, for example, are now sufficiently well understood that they would presumptively fall outside of our analysis.[31]

This final point is important, even if one disagrees with our proposal. Though self-driving cars and large language models (LLMs) are two prominent deployments of AI that are the focus of current policy debates, future developments in AI may rely on novel architectures. Thus, a framework that is too narrowly focused on deep learning or LLMs may ultimately fail to be future-proof. Moreover, if AI technology does advance, it may be the case that as

---

[28] *See infra* Section III.B.ii.

[29] It was at one point incredibly important to work out the liability implications of automobiles and standards of safe driving likely encountered a period of flux as the use of the new technology expanded. But society caught up and defendants today are unlikely to be able to defend lawsuits based on safe driving by claiming that the reasonable person of ordinary prudence lacks knowledge of how to drive safely. This is not to say the spade work of working out said standard was not important (it was!) but rather to suggest that in areas of rapid technological advancement it can be perilous to focus on the technology of today that may be displaced tomorrow by a new paradigm. *See* Kyle Graham, *Of Frightened Horses and Autonomous Vehicles: Tort Law and its Assimilation of Innovations*, 52 SANTA CLARA L. REV. 1241, 1242 (2012) (discussing five ways in which the tort system assimilates new innovations); Gregory N. Mandel, *History Lessons for a General Theory of Law and Technology*, 8 MINN. J.L. SCI. & TECH. 551, 552 (2007) (providing lessons from history about how law incorporates new technologies).

[30] For example, what warnings product manufacturers need to provide to end users depend, in part, on which accidents are foreseeable. *See infra* Section I.A. In an environment in which AI/human interactions are novel, this can raise difficult questions. If LLM development stalls out and interactions with LLMs become normal, the problem of unexpected accidents will potentially vanish. If, however, we expect AI development to continue to encompass new model type with more power than questions about the transition are likely to recur. *See* Ram Iyer, *Meta's Chief AI Scientist Yann LeCun Reportedly Plans to Leave to Build His Own Startup*, TECHCRUNCH (Nov. 11, 2025, 6:58 AM PST), https://techcrunch.com/2025/11/11/metas-chief-ai-scientist-yann-lecun-reportedly-plans-to-leave-to-build-his-own-startup/ (noting that LeCun is skeptical that LLMs can achieve general intelligence and plans to form a start-up to develop "world models" which are "AI system[s] that develops an internal understanding of [their] environment so [they] can simulate cause-and-effect scenarios to predict outcomes").

[31] At minimum, we think they fall outside of the scope of our analysis when applied to specialists such as lawyers who are on at least inquiry notice of the issue and who should be expected to understand the limitations of these tools before deploying them. *See infra* Section II.B.

society normalizes a particular technology, new technologies again test the limits of foreseeability. Today, for example, hallucinations may be sufficiently metabolized to be handled with traditional legal tools but new systems or the discovery of new failure types may start the cycle anew. We thus aim to develop a framework that is technology-agnostic and to then apply it to current examples of AI accidents rather than focus more narrowly on the problems of current technologies.

Our Article proceeds to make this case in four Parts. The first discusses previous work on AI liability and the general effort to find a category in which to place AI liability. Here we also discuss what we see as the shortcomings of these approaches. The second then lays out our framework at a high level and connects it to some of our theoretical inspirations. In doing so, we also discuss traditional tort law liability doctrines and show how their contours broadly fit the model that we are advancing for dealing with the problems of AI. Third, we provide a taxonomy of known mechanisms that cause AI errors drawing on the extensive computer science literature on the topic. Fourth and finally, we provide a proposal for liability that builds on this taxonomy and discuss its benefits.

I. PREVIOUS PROPOSALS (A SAMPLE)

Our core view is that liability for the actions of artificial intelligence systems should be based on fault. Accordingly, we find much to agree with when we consider applications of traditional approaches to liability to this new technology. That's a good thing. Radical revisions to old bodies of law are ideally accompanied by radical arguments either in theory or based on the facts of the case. Accordingly, this Section discusses proposals to apply traditional legal *categories* to the problems of artificial intelligence to both highlight points of agreement and to identify ways in which we believe the *categorical* approach to AI liability gets things wrong on either a conceptual level or is underdetermined as a matter of application.

For our purposes, we are most interested in how these approaches address the foreseeability gap. In essence it appears to us that there are three main strategies. One, of course, is to leave the gap open and accept that odd errors are so unpredictable that losses lie where they will. Understandably this approach is unpopular in the literature as scholars writing in this space are doing so often to avoid such a conclusion. Second is to attempt to narrow the gap by reworking traditional doctrines to argue that they sweep in many potential concerns. Third, we can adopt some form of strict liability that functions as a fallback rule to engage in loss spreading where traditional moves fail.

The remainder of this Part attempts to cover a sample of the approaches that have been offered to adapt tort law to regulate AI. We acknowledge that this list is underinclusive in two ways. First, we are not able to cover all the potential

doctrinal adaptations.[32] Second, scholars have proposed many different approaches to the categories below and we are unable to canvas all the various twists on these ideas. We believe that our general skepticism articulated below can be responsive to some of these other scholars, but we appreciate that we cannot be responsive to every proposal in this space.[33]

A. Strict (Products) Liability

An initial approach would be to expand the concept of strict product liability to these new artificial intelligence systems. This approach has a lot of appeal and, in some ways, overlaps with our treatment of the problem. In particular, the three main categories of product liability claims—failure to warn, manufacturing defect, and design defect—each offer some means of addressing issues posed by artificial intelligence systems. Responding to the appeal of this approach, several legislative proposals have looked to apply products liability concepts to regulate AI developers and to draw the line between developer and "deployer" liability.[34]

First, failure-to-warn claims seem attractive where there is an information asymmetry between the manufacturer and the end user.[35] Second, manufacturing defect claims may also be promising where the machine fails to be trained properly or up to the specifications of the designer.[36] Third, and most interesting, are design defect claims—those asserting that the product itself was

---

[32] Notably, we do not discuss the analogy to wild animals, though our discussion of tort law in *infra* Section II.A we think covers our skepticism of this analogy. We also decline to discuss procedural adaptations that might address some of the evidential and procedural issues present in AI litigation.

[33] We note that some of the scholarship we do not respond to is explicitly focused on the problem of AVs. Given the importance of automobile accidents to tort law and the surge of interest in AVs over the past decade this makes sense. That said, because we are interested in pushing towards a solution that is responsive to AI in general, we wish to avoid the weeds of this issue.

[34] *See* Frank Young, *AI as a Product: The Next Frontier in Product Liability Law*, U. ILL. CHI. (Oct. 13, 2025), https://library.law.uic.edu/news-stories/ai-as-a-product-the-next-frontier-in-product-liability-law/ (discussing the AI LEAD Act introduced in the Senate); *see also* Evangelos Razis, *Deployers and Developers: Colorado's Partnership for AI Governance*, IAPP (Oct. 8, 2024) (discussing how the Colorado AI Act regulates developers and deployers of AI systems for "consequential decisions").

[35] RESTATEMENT (THIRD), *supra* note 24, § 2(c) (holding manufacturers liable where the "foreseeable risks of harm posed by the product could have been reduced or avoided by the provision of reasonable instructions or warnings by the seller or other distributor, or a predecessor in the commercial chain of distribution, and the omission of the instructions or warnings renders the product not reasonably safe"); *see also id.* § 10 (requiring manufacturers post-sale warnings where, among other things, "the product poses a substantial risk of harm to persons or property" and "those to whom a warning might be provided can be identified and can reasonably be assumed to be unaware of the risk of harm").

[36] *Id.* § 2(a) (holding manufacturers liable for a "manufacturing defect when the product departs from its intended design even though all possible care was exercised in the preparation and marketing of the product").

negligently designed and that an alternative design would have been safer while preserving the product's utility.[37]

At a high level, these are three categories of wrongs that a developer of an AI system might commit, and the law should offer a remedy. The issue, however, is that stated at this level of generality, the claims leave much to be determined. What sort of errors should an AI system developer warn end users about?[38] The restatement requires that the harms be foreseeable, but application of this criterion is nontrivial in light of both the opacity of AI systems and the novelty of AI/human interactions. Similarly, what exactly constitutes a manufacturing defect in an AI model?

Perhaps the most important and most difficult question is how to apply design defect claims to AI systems. Deep neural networks, in essence, use a network of parameters to map inputs to outputs.[39] While developers of these systems make a variety of choices about high level questions such as the structure of the system, its hyperparameters, and the data that is used to train the model, many errors the model makes will be emergent and not a direct consequence of any of these choices in an obvious way.[40] In certain trivial cases, these design choices might be so irresponsible so as to render the model obviously defective, but the hard questions of liability in this space arise in cases where the system seems to function fairly well most of the time. Indeed, the promise of AI systems (and their appeal to users) is precisely that in many cases they can outperform many humans along some important dimension.[41]

Moreover, to the extent the modern trend is to require a reasonable alternative design, vexing questions arise of how to assess proposed redesigns of such poorly understood systems.[42] Again, some simple cases may present themselves. For example, a viral rumor spread that OpenAI's models would no

---

[37] *Id.* § 2(b).

[38] *See infra* Section IV.B (discussing questions of notice within our framework).

[39] *See* JOHN D. KELLEHER, DEEP LEARNING 64-99 (2019) (describing the basic foundations of deep learning algorithms). *See generally* ANIL ANANTHASWAMY, WHY MACHINES LEARN: THE ELEGANT MATH BEHIND MODERN AI (2024) (providing a history of AI development as well as a mathematical overview of key concepts).

[40] Briefly, while a neural net is fundamentally a set of simple functions linked together one can link the neurons together in various ways to perform certain tasks better. Thus, for example, early innovators found that convolutional neural networks (CNNs) could be used for image recognition. ANANTHASWAMY, *supra* note 39, at 360-81. Moreover, when designing a CNN, one has to make a variety of choices about its structure such as the number of layers of particular types that cannot be learned and are thus termed the hyperparameters of the model. *Id.*

[41] *See* Shavell, *supra* note 2, at 243-44 & n.2 (observing that AVs have the potential to save large numbers of lives via the reduction of accidents due to human error though noting some potential reasons for caution about the magnitude of the effect).

[42] RESTATEMENT (THIRD), *supra* note 24, § 2(b).

longer provide medical and legal advice.[43] Though this rumor was not true (and OpenAI actually launched a health product), one can imagine a situation in which such a prophylactic design choice would make sense.[44] Still, to the extent the concern is with the emergent behavior of these products, it is not obvious how to weigh the risk and utility of an alternative design that would also have its own emergent behaviors.

In an early and influential article, David Vladeck points out the weakness of products liability for hard cases of AI liability.[45] Vladeck considers the following hypothetical:

> [S]uppose daredevil Mrs. Arnold again darts in front of a driver-less vehicle. In this scenario, assume that the vehicle's designers have programmed instructions that it should at all costs avoid striking a human. And further assume that, in order to avoid hitting Mrs. Arnold, the vehicle has only two choices: It can either (1) take evasive action by making a sharp turn that would likely avoid any impact with Mrs. Arnold, but would also likely result in the vehicle striking a brick wall (thereby risking the safety of the vehicle's occupants), or (2) apply the brakes with great force, even though, given the distance and vehicle's speed, that course of action would not guarantee that the vehicle would stop before hitting Mrs. Arnold; nor would it ensure that there would be no collision between it and the car following closely behind.[46]

Vladeck imagines that the car takes none of the above routes and indeed disobeys its maker's instructions to run into the brick wall and trust the safety of the vehicle and instead attempts to decrease the car's speed and contacts Mrs. Arnold at a lower speed.[47] He notes that a manufacturing defect claim won't work as long as it is premised on the mechanics of the car working properly.[48] Nor does it seem

---

[43] Emma Roth, *No, ChatGPT Hasn't Added a Ban on Giving Legal and Health Advice*, VERGE (Nov. 3, 2025, 5:20 PM EST), https://www.theverge.com/news/812848/chatgpt-legal-medical-advice-rumor.

[44] *Introducing ChatGPT Health*, OPENAI (Jan. 7, 2026), https://openai.com/index/introducing-chatgpt-health/.

[45] Vladeck, *supra* note 2, at 143-44.

[46] *Id.* at 144.

[47] *Id.*

[48] *Id.* at 145. One could quibble here that the car is indeed not up to spec in the sense that it has disobeyed its instructions and thus malfunctioned. This argument, however, raises a level of generality concern: the product has been, we assume, built to its intended design in the sense that the system has rolled off the factory line as intended yet it displays emergent behavior that is not what its creators intended on some other level. Again, the analogy to scholastic degrees seems relevant. The degree is intended to both convey that the recipient has completed the relevant coursework at an appropriate level *and* that the recipient indeed has the relevant knowledge that one would expect of someone with said credential. Manufacturing defect law

that a design defect claim will work under either a consumer expectations theory—consumers have no reasoned expectations about what the car ought to have done—or a risk-utility theory that, as discussed, requires an alternative design that plaintiff is unlikely to be able to develop.[49]

This discussion of strict liability also to this point does not consider true strict liability as exists for abnormally dangerous activities. Under either regime, however, there is also a concern with too aggressive an approach to strict liability.[50] AI systems' potential to cause harm is large—indeed, this is what prompts the desire for liability. However, if strict liability is applied in a way that could lead to massive losses for developers when a lawsuit succeeds, it may paradoxically lead to underinvestment in safety.[51] If any successful lawsuit could destroy a developer financially, there is little incentive to take precautions.

B. Enterprise Liability

Another idea, suggested by Vladeck, is to hold the entire chain of users—from the developer of the system to the end user—liable as a common enterprise. Vladeck observes that the trend in artificial intelligence is away from systems that are tools or mechanisms deployed by humans to accomplish some end and towards machines that are deployed into the field to act independently.[52] In particular, they will "act independently of direct human instruction, based on information the machine itself acquires and analyzes, and will often make highly consequential decisions in circumstances that may not be anticipated by, let alone directly addressed by, the machine's creators."[53] The resulting robots look

---

seems to focus on the former, formal aspect of making a good—that it is an appropriate exemplar of its species—and not on the latter, substantive aspect—that the product as designed is good for what it says on the tin. *See supra* notes 35-37 and accompanying text.

[49] Vladeck, *supra* note 2, at 145. Indeed, at the risk of being glib, if plaintiffs and his or her lawyers can develop such an alternative design, they would be perhaps well advised to commercialize it to recoup the damages to Mrs. Arnold in some other way.

[50] In the context of AVs, Steven Shavell has argued that the economically optimal liability rule is not conventional strict liability for manufacturers or negligence but rather strict liability for manufacturers with *damages paid to the state*. Shavell, *supra* note 2, at 244-47.

[51] This is a familiar point in discussions of strict liability which only allows firms to escape tort judgments via insolvency whereas a negligence standard allows them to limit liability by adhering to the relevant standard of care. An interesting proposal to address this issue comes from Gabriel Weil who argues that courts or legislatures should impose punitive damages in cases that are associated with uninsurable risks but which have compensable harms small enough that punitive damages can be awarded without driving the defendant into bankruptcy. Gabriel Weil, *Tort Law Should Be the Centerpiece of AI Governance*, LAWFARE (Aug. 6, 2024 10:03 AM), https://www.lawfaremedia.org/article/tort-law-should-be-the-centerpiece-of-ai-governance. This proposal "pulls forward" the costs of catastrophic risk by attempting to force developers to internalize the costs of generating catastrophic risk in "near miss" cases. *Id.*

[52] Vladeck, *supra* note 2, at 121.

[53] *Id.*

more like persons than they do machines and their creators resemble teachers rather than users of a product. And while Vladeck notes, as we do, that some cases will still be covered by traditional product liability, others will involve cases in which the machine's ability to "think" causes it to disobey the rules that it has been given.[54]

Looking to the problem of truly unforeseeable errors, Vladeck is unwilling to let losses lie where they will and also to claim that traditional doctrines will suffice to cover the foreseeability gap.[55] His approach is to then argue that a system of strict liability ought fill the gap.[56] Interestingly, Vladeck admits that this is, in essence, an insurance approach that is implemented via liability.[57] Unlike proponents of alternative schemes who argue that fault-based (or fault-adjacent) principles will solve the problem, Vladeck accepts that the errors in question will be truly unforeseeable.[58] Where he differs from other approaches is that he observes that a strict liability approach that holds only the manufacturer (Vladeck has in mind the maker of an AV) liable may fail to create proper incentives for safety or to engage in appropriate cost spreading if third-party radar or computer code is that actual cause of the failure.[59] His solution is apply a variation of "common enterprise" liability to hold all entities involved in producing the vehicle jointly and severally liable.[60] The parties can then bargain for indemnity and thus distribute their new responsibilities as insurers among themselves.[61]

---

[54] Vladeck, *supra* note 2, at 123, 127 ("The first is how to apply the law of products liability on the assumption that any liability concern with the machine is the result of human (but not driver) error—that is, a design or manufacturing defect, an information defect, or a failure to instruct humans on the safe and appropriate use of the product. In my view, the application of these reasonably settled principles is a straightforward one, and there is no justification for treating even autonomous thinking machines differently than any other machine or tool a human may use, except, perhaps, holding them to a higher standard of care.").

[55] *Id.* at 146.

[56] *Id.* ("In this instance, a system of strict liability would, in effect, impose a court-compelled insurance regime to address the inadequacy of tort law to resolve questions of liability that may push beyond the frontiers of science and technology.").

[57] *Id.*

[58] *Id.* at 149 ("That principle could be engrafted onto a new, strict liability regime to address the harms that may be visited on humans by intelligent autonomous machines when it is impossible or impracticable to assign fault to a specific person.").

[59] *Id.* at 148-49.

[60] *Id.* at 149.

[61] *Id.* One puzzle here is why the common enterprise liability is needed to reach this solution. If a manufacturer of an autonomous vehicle is concerned that advanced parts or computer code will fail, they can seek out some form of indemnity when bargaining for the equipment and potentially use willingness to provide such assurances as a relevant factor when deciding who to contract with. In other words, Coasean bargaining should still take place among sophisticated parties in long-term repeat relationships. Vladeck suggests that "If it is in fact impossible to identify the cause of the accident, then the manufacturer would likely have no reasonable grounds for an indemnity or contribution action, and would thus be saddled with the entire

Vladeck's approach is important and provocative.[62] Still, we believe that something more can be done to close the foreseeability gap, especially as these systems have become better understood and new applications of machine learning have become more popular. To the extent our approach (or any of the other approaches) fails to solve the foreseeability problem in a way that is satisfactory, Vladeck's approach to strict liability is one potential way to fill the gap.

C. Employed Algorithms and Respondeat Superior

As discussed in the introduction, we are particularly interested in cases where an employer uses an AI system as an employee or as an advanced tool within a workflow. This issue has become more pressing as leading AI companies push into what they term "agentic models," capable of executing complex workflows.[63] Companies believe these advanced models can serve as digital employees, replacing the work of human individuals.[64] The question then becomes how to hold companies liable for torts or other accidents committed by these employed algorithms.

Mihailis Diamantis suggests a straightforward solution: treat these algorithms as employees and apply traditional doctrines of employer liability.[65] In particular, he proposes to adapt *respondeat superior* to algorithms by making two adaptations to the doctrine. The first is to recognize that employers can employ algorithms and the second is to adapt the two prongs of the doctrine (the benefit prong and the scope of employment prong) to algorithms.[66] Diamantis does so by adopting what he terms the "beneficial-control test" in which a corporation is liable for harms due to algorithms when it "claims substantial benefits from the algorithm's operation and exercises substantial control over it."[67]

As with product liability, we agree that there is something to this suggestion. It would make little sense to allow employers to avoid liability simply by

judgment." *Id.* at 128. But again, it's not clear why the manufacturer and component manufacturers cannot design a more elaborate system. Changing to a common enterprise baseline may alter the distributional impacts of strict liability or, perhaps, reduce transaction costs by spurring bargaining but Vladeck's article does not make such a claim.

[62] Vladeck also suggests that at a future date a court could treat the vehicles as agents with a form of legal personhood. *Id.* at 149-50.

[63] *See, e.g.*, *Agents*, OPENAI, https://platform.openai.com/docs/guides/agents (last accessed Dec. 2, 2025).

[64] *See What is a Digital Worker*, IBM, https://www.ibm.com/think/topics/digital-worker (last accessed Dec. 2, 2025)

[65] Diamantis, *Employed Algorithms*, *supra* note 2, at 804. Lior also makes the analogy to *respondeat superior* but seems to argue that this will result in a form of strict liability given that an AI agent always acts within the scope of its employment. Lior, *supra* note 2, at 1096-1100.

[66] *Id.* at 844, 847-48.

[67] *Id* at 848. Diamantis notes that both elements may involve multifaceted inquiries. *Id.*

substituting an algorithm for a human employee. Indeed, as Ryan Abott observes in the context of tax law, to hold otherwise would function to subsidize automation artificially.[68] And Diamantis notes that the law encountered a structurally similar problem when employers attempted to avoid liability via the use of contract workers and thus to argue that the more limited doctrines of attribution available for harms committed by independent contractor shielded them from liability.[69] If a human employee discriminates in hiring and the employer is liable, it seems equally sensible that an employer who sloppily uses an AI system to discriminate in the same way should be liable under similar principles. This approach also aligns with the attractive notion that regulation should be technology-neutral—neither advantaging nor disadvantaging computerized systems.

This is, however, where the concern that animates this paper comes in. In many ways the law governing vicarious liability or direct liability of employers or the torts or other action of their employees is based around the assumption that employers know the ways in which typical employees will err. This concern both motivates the fairness of seemingly strict doctrines such as *respondeat superior* and links the doctrine to its deterrence function.[70]

The problem artificial intelligences pose for this paradigm is not that they err too, nor is it that they err more than human employees—indeed the appeal of some of these systems (think of self-driving cars) is that they have lower error rates than human actors. The trouble rather is that these artificial intelligence systems will err in different ways than human employees and thus managers applying the expectations of human employees to managing their digital employees may find that they're not effectively able to catch and prevent errors using their conventional processes.

To what degree this critique applies to Diamantis's proposal depends on the application of his control prong. He suggests that evaluation of control would "include the power to design the algorithm, terminate its operation, modify it, monitor it, and override it" and that "the more powers a corporation has, the more control it has."[71] If liability is triggered only at high levels of control then the expansion of foreseeability is relatively modest and the gap between our approach and Diamantis's is relatively small.[72] At lower levels of control,

---

[68] ABBOTT, *supra* note 2, at 4-7.

[69] Diamantis, *Employed Algorithms*, *supra* note 2, at 830-40 (providing a history of employer efforts to limit liability via the use of contract workers); *see* RESTATEMENT (SECOND) OF TORTS § 409 (AM. L. INST. 1965) (setting out the default rule that "the employer of an independent contractor is not liable for physical harm caused to another by an act or omission of the contractor or his servants").

[70] *See infra* Section II.A.

[71] Diamantis, *Employed Algorithms*, *supra* note 2, at 848.

[72] In the extreme case the "employer" has designed the algorithm, can modify it, and can override it. We might still insist on traceability of the error to a known mechanism of failure, but

however, if one accepts our claims about the unforeseeability of some AI errors, the model pushes towards asking the employer of the algorithm to internalize unforeseeable and hard to control errors. Moreover, one will have to ask about how to address digital employees built off a general model made by an AI developer but deployed by some other company (though Diamantis notes that his model contemplates the potential for joint employment).[73]

Over time it's possible that societal norms will shift in such a way that managers learn how to manage these digital employees and this problem becomes a nonissue. In some cases the harm done by the digital employee at the direction of the employer will so closely resemble the type of harm done by a human employee that considerations of fairness, deterrence, and parity will all push in favor of the solution Diamantis proposes. Once more, we think there will be cases in which the foreseeability gap will frustrate this approach or at least raise questions about its applicability.

D. Negligence

Another proposal is to apply traditional negligence analysis. Developers would be required to act with ordinary prudence when designing AI systems, and users would similarly be expected to exercise reasonable care when using them. At a high level, we agree with this fault-based approach, and the solution developed below reflects this orientation. Nevertheless, a high-level invocation of negligence principles stumbles on two conceptual issues that our proposal seeks to address.

First, it leaves undefined what constitutes ordinary care in developing AI models. As news reports and research show, AI's stochastic and unpredictable nature makes it quite unlike traditional products—even other software, which follows deterministic patterns.[74] It's true that traditional software can also err and at times these errors can be large in magnitude—consider the massive losses caused by defective trading algorithms—their deterministic nature makes them complicated but amenable to audits.[75] The stochastic nature of many artificial intelligence systems and the barriers to explainability they present makes them

---

at such a high level of control and creation the gap may not be that large or alternatively the two approaches can merge. In either case, we agree the question is about whether to attribute liability to the controller of the algorithm and in such a case fairness might dictate leaning on the side of the controller.

[73] Diamantis, *Employed Algorithms*, *supra* note 2, at 849.

[74] *See* Schneier & Sanders, *supra* note 9; *see also* Vladeck, *supra* note 2, at 123.

[75] *See, e.g.*, Nathaniel Popper, *Knight Capital Says Trading Glitch Cost It $440 Million*, N.Y. TIMES (Aug. 12, 2012, 9:07 AM), https://archive.nytimes.com/dealbook.nytimes.com/2012/08/02/knight-capital-says-trading-mishap-cost-it-440-million/ ("The problem on Wednesday led the firm's computers to rapidly buy and sell millions of shares in over a hundred stocks for about 45 minutes after the markets opened.").

more akin to human beings whose intentions are similarly difficult to truly probe.[76] Thus it remains unclear how to assess negligence when developing or deploying an AI system. To be sure, we can imagine easy case of clear negligence in designing or deploying an AI system that would make for an easy case, but the harder cases are the ones of interest.

One way to address questions of ordinary care is to adopt a professional care standard as Bryan Choi and Chinmayi Sharma.[77] Choi builds on previous work to suggest that professional care is appropriate when an occupation is an inexact art that requires subjective judgments that carry the risk of large harms yet fulfills a socially important function.[78] Applying these factors to the work of AI development, Choi reaches an uncertain conclusion about whether they qualify for a professional care standard under his approach. Still, he identifies professional care as one way to move away from questions about what is objectively reasonable in AI design. Sharma goes further and proposes formal professionalization in which institutions would regulate AI engineering via standardized training and licensing.[79]

Both suggestions are valuable, but both also present problems. First, in a rapidly evolving field it is not obvious that professional custom exists in a sufficiently definitive form to provide much guidance in this area. Certainly, some baseline standards may exist, but unlike medicine and law AI engineering is potentially too young for professional standards to do too much work at this juncture outside of easy cases of malpractice. Second, to the extent professional custom does come to exist, as Choi recognizes, this is a different question from whether the law should defer absolutely to the customary standard. The traditional rule, rather, is that custom is probative of ordinary care but not dispositive.[80] In this sense, we agree with Choi's tentativeness about this suggestion: professional custom may help with some cases and may indeed

---

[76] Indeed, foundational questions about why deep neural networks generalize from their training data to solve new problems effectively are unanswered and remain the subject of active research. *See* SIMON J.D. PRINCE, UNDERSTANDING DEEP LEARNING 402-20 (2025), *available at* https://udlbook.github.io/udlbook/ ("Many questions remain unanswered. We do not currently have any prescriptive theory that will allow us to predict the circumstances in which training and generalization will succeed or fail. We do not know the limits of learning in deep networks or whether much more efficient models are possible. We do not know if there are parameters that would generalize better within the same model. The study of deep learning is still driven by empirical demonstrations. These are undeniably impressive, but they are not yet matched by our understanding of deep learning mechanisms.").

[77] Choi, *supra* note 2; Sharma, *supra* note 2.

[78] Choi, *supra* note 2, at 306.

[79] Sharma, *supra* note 2 at 1105-09.

[80] RESTATEMENT (SECOND), *supra* note 69, § 295A cmt. c ("Any such custom is, however, not necessarily conclusive as to whether the actor, by conforming to it, has exercised the care of a reasonable man under the circumstances, or by departing from it has failed to exercise such care. Customs which are entirely reasonable under the ordinary circumstances which give rise to them may become quite unreasonable in the light of a single fact in the particular case.").

provide evidence of (or against) negligence, but its applicability is uncertain and limited. As for professionalization, Sharma's suggestion is provocative and perhaps reflects a desirable end-state of affairs under certain assumptions. We focus, however, on a (somewhat) orthogonal question of what duties ought to govern in the meantime.

Second and more importantly (and as we'll discuss later), even if questions about the standard of care can be worked out, critical parts of negligence law are designed around a notion of fault and foreseeability that asks us to consider the likely consequence of our actions.[81] Critically, however, this assumption allows potential *plaintiffs* to structure their affairs also based on common notions of foreseeability. Negligence law will sometimes hold people liable for failure to supervise carefully but also will not hold us liable for accidents that would be unforeseeable to the ordinary person.[82] But this raises the question of what *types* of accidents are foreseeable and *who* can foresee them.[83]

Anat Lior has provided an interesting perspective on these foreseeability issues. She points out that tort law leaves foreseeability questions to the factfinder and thus implicitly leaves open the question of the level of generality to be applied.[84] The result is flexibility to consider whether the foreseeable issue is whether to release a product that one knows will *predictably fail* in unforeseeable ways or instead to focus on whether the particular failures are *unpredictable failures*.[85] That suggestion seems right to us in theory but also would seem to push in the direction of either pure products liability of the former is adopted or towards some undefined duty of care under the latter approach.

Elements of negligence (and particularly breach and proximate cause) thus pose difficult questions about foreseeability when applied to artificial intelligence. It's not that negligence law can't handle these questions. Indeed, it's proven to be a flexible and adaptable doctrine as technology has advanced. Rather, what is needed is a theory that goes one level deeper into explaining how these foreseeability questions will actually work when applied to new systems.

With respect to breach, in a recent article Diamantis has offered one potential solution to the problem. Diamantis notices, as we do, that AIs do not "think" like humans do. This frustrates attempts to compare humans to AIs and thus to use a human baseline for a standard of breach.[86] Similarly, he argues that

---

[81] *See infra* Section II.A.

[82] *See infra* Section II.A.

[83] Per the example in the introduction, do ordinary people foresee AI hallucinations? Do ordinary lawyers? Do heightened standards of professional competence change the calculus? Certainly, over time social expectations and understandings evolve to make certain accidents more or less foreseeable and will alter the shape of liability. But there is no guarantee that the frontier of AI model design will remain fixed.

[84] Lior, *supra* note 11, at *4

[85] *Id*.

[86] Diamantis, *Reasonable AI*, *supra* note 2, at 596-600.

comparing AI to AI creates baseline problems: Is an AI that drives better than humans but worse than the state of the art unreasonable?[87] Perhaps more important, Diamantis questions whether it will be possible to administer a reasonable AI baseline given the small number of AI systems.[88] To solve these problems, he proposes a standard that defines objective reasonableness in terms of a weighted average of harm among all actors, human and AI.[89] An AI that fails to meet this test is not reasonable and, if it causes harm, is negligent.[90]

This is an interesting and compelling proposal, though the scoping of the task seems potentially difficult when we move from specialized models (Diamantis motivates his analysis using the example of self-driving cars) to more general-purpose models. Still, we think further analysis about foreseeability is needed. This is mainly because any definition of harm "caused" by the AI will sweep in a question of proximate cause which then brings in foreseeability concerns. Is crashing into a jet on a tarmac part of the unit of harm? It depends on the scope. Thus, even if we accept Diamantis's objective standard, we think more is needed to define what exactly AI is "responsible" for.

E. No Fault/No Liability Approaches

The final category of proposals would absolve AI developers of fault and instead introduce a no-fault compensation scheme, similar to workers' compensation.[91] At the heart of such a suggestion is really two questions: (1) is litigation well suited to assessing who bears the costs of AI errors and (2) if not, what scheme of compensation, if any should fill the gap? Because we are concerned with the first question (and believe the answer is yes) we discuss that part of the question here.

There seem to us to be, broadly speaking, two ways that proponents of this approach reach their conclusion. The first is a broadly economic concern that state regulation, whether in the form of bills or perhaps in the form of the common law, will stifle innovation through some combination of ill-considered requirements and high compliance costs due to the need to manage a large number of state standards.[92] Dean Ball and Alan Rozenshtein, for example, argue that Congress should pre-empt state bills in this area on the ground that AI policy is of national importance and therefore should not be subject to one

---

[87] *Id.* at 600. *But see* ABBOTT, *supra* note 2, at 58-60 (defending a "reasonable robot" standard).
[88] Diamantis, *Reasonable AI*, *supra* note 2, at 600-01.
[89] *Id.* at 602-04.
[90] *Id.*
[91] Vladeck, *supra* note 2, at 149 n.95.
[92] *See* Dean W. Ball & Alan Z. Rozenshtein, *Congress Should Preempt State AI Safety Legislation*, LAWFARE (June 17, 2024, 2:00 PM), https://www.lawfaremedia.org/article/congress-should-preempt-state-ai-safety-legislation (making a version of this argument with respect to state AI bills).

state's view of the risk-benefit calculus that may not be appropriate for the national economy.[93] Though they focus on state statutory law the same could be said for state negligence law (or any other branch of common law liability). Though state common law could converge over time, in the short-term chaotic results could lead to both of the same problems.

Though, in theory, a scheme of federal regulation could include a private right of action, but the force of this argument seems to suggest relatively minimal liability for manufacturers of AI system. Ball and Rozenshtein, for example, disclaim interest in a broad, "Section 230-like liability immunity for model developers" but conclude that "when the industry is still nascent, a rebuttable presumption of user responsibility likely strikes the right balance between innovation and accountability."[94] Any broad and flexible private right of action seems destined to lead to protracted litigation and some number of inconsistent judicial opinions. While the Supreme Court can step in to resolve splits in a national way (unlike with state tort law) a broad cause of action would still likely require a number of cases to give shape to its relevant requirements.[95] A concern for innovation and the need to limit complexity seems to leave little room for court adjudications of liability.

While this account focuses on innovation, a second theory is possible. In particular, the foreseeability gap is the result of applying traditional tort principles to a novel setting. Because those principles are being applied to an area that presents unusual risks, the accidents that result are unforeseeable and seem to be the responsibility of no party. If we believe in the integrity of the traditional rules and their ability to capture important aspects of interpersonal morality, then perhaps there is no real problem. If the accidents are truly unpredictable and unusual then perhaps no one is morally or legally responsible for them. We could consider whether to ask the makers of AI models to abstain from the actions all together, but if we judge that AI development is a socially valuable activity, then that approach will not do.

In either case, we then return to the second question of how to compensate those who lose out due to the unforeseeable accidents of AI. The answer would seem to be some form of insurance be that a program of social insurance, a private market for accident insurance, or a court-imposed system as Vladeck

---

[93] *Id.*

[94] *Id.*

[95] Constitutional guarantees and the Court's antitrust jurisprudence illustrate this dynamic. Even in the area of FDA pre-emption where much of state law is pre-empted, the Court has still needed to step in multiple times to clarify the scope of pre-emption. *See* Mitchell C. Johnston, Marcia M. Boumil & Gregory Curfman, *A New Supreme Court Ruling on Drug Liability*, 332 JAMA 607, 607 (2019) (discussing a 2019 Supreme Court decision (where the complaint was filed in 2010) that resolved a question left open by a 2009 decision of the Court).

suggests.[96] Whether or not there is a political appetite for such a solution remains to be seen. Perhaps in some cases the gains from AI use in, say, autonomous vehicles will be so great the companies will accede to a form of de facto strict liability that they will offset with insurance. In other cases, the picture may not be so clear.

But even if we ignore practical problems, we are unconvinced that a broad reliance on regulation and insurance to replace tort liability is attractive. Artificial intelligence researchers have long thought about the problem of alignment and how to ensure these systems engage in pro-human and pro-social behavior. It seems sensible as with other areas of developing technology for the law to play some role in encouraging people to put out products that are safe for human use. This is especially true, we think, where companies are not only putting out these products to interact with direct users where we might think that principles of contract law would apply, but also where they are explicitly encouraging those users to deploy their models to interact with third parties (who might reasonably expect their interaction with the user to not be mediated by AI). Someone who is discriminated against in a job search because an employer has purchased an AI model has not consented to an interaction with an artificial intelligence and indeed might have a justified view that his or her application will be considered by a human.[97] We also think it's sensible that someone should take some care to make sure that the interaction is safe and complies with applicable law rather than foisting losses onto the recipient of the treatment. As with the complexity objection our view is that this issue can be overcome, and the law can sensibly hold parties responsible for AI errors.

* * *

In canvassing this sample of approaches to the problem of expanding tort law to regulate artificial intelligence we can note a few key themes. The first is that depending on the use case, artificial intelligence systems sometimes resemble traditional categories in tort law such as products or employees. In such cases it seems natural to apply these traditional categories. The second is that many

---

[96] *See* Anat Lior, *Insuring AI: The Role of Insurance in Artificial Intelligence Regulation*, 35 HARV. J.L. & TECH. 467, 471-74 (2022) (providing an examination of how insurance law applies (and should apply) to AI and the advantages and disadvantages of AI insurance); *see also* Jin Yoshikawa, Note *Sharing the Costs of Artificial Intelligence: Universal No-Fault Social Insurance for Personal Injuries*, 21 VAND. J. ENT. & TECH. L. 1155 (2019) (arguing in favor of a no-fault scheme of social insurance).

[97] Nor is the presence of a human decision maker immune to this type of objection. One of us recently encountered a story by an acquaintance who was encouraged to apply for a role in the recruiter's company but discovered that after applying their resume was screened out by a team mechanically employing various minimum qualification tests that the recruiter thought were not applicable in this context. The acquaintance was aggrieved, in part, because the initial representations had led them to believe that they would not be screened out on this basis.

commentators acknowledge that, in addition, a key problem for the regulation of AI systems is their unpredictability which makes them hard to fit into a traditional box—what we term the foreseeability gap. Third and finally, there are two general approaches that have been taken to the problem. One is to close or narrow the gap by either fitting a traditional doctrinal framework onto the problem or adjusting levels of generality such that AI errors are included into the scope of the doctrine. The second is to admit that the gap presents a problem that is a poor fit for tort law and to allow it to go unfilled or to fill it with something that resembles a form of insurance. The remainder of this Article turns to a different approach to addressing the gap.

II. TORT LAW AND THE PROBLEMS OF AI

A. The Human-Fallibility Assumptions of Tort Law

In Part I, we discussed how various proposals for AI liability struggle with the foreseeability gap. They do so because tort law (and many other areas of law) incorporates what we have called human-fallibility assumptions. Namely the law allocates responsibility based on assumptions about how and when humans err and social knowledge of these assumptions. Our contention in this Article has been that artificial intelligence systems undercut the applicability of this assumption because of the strange way in which they are prone to error. Still, before making this claim, we have to defend the first part which is that tort law embeds this assumption at all. This Section therefore walks through several doctrines we believe demonstrate this basic assumption.

At the start, we note that this is actually not as exotic an argument as it may appear. The duty of ordinary care holds individuals to an objective standard. When considering questions of breach, typically individuals with diminished capacities cannot rely on those diminished capacities to avoid liability. The result of this rule is that tort law enables us to generally make the assumption that others will act with reasonable care towards us as we structure our affairs and that we will be protected if they fail to do so. Moreover, the ability to practically apply this assumption to our lives then relies on *our* judgments as potential plaintiffs about what it means for typical people to take ordinary care. Thus, when I'm driving on the highway, I can generally assume that most people are also driving carefully. I can also take precautions based on the normal lapses of other people in order to make myself additionally safer.[98] To be sure, tort law

[98] Indeed, one traditional justification for negligence law over strict liability is that it encourages me to take cost justified precautions to protect against other people's negligence if I am the least-cost avoider of the accident. *See* Richard A. Posner, *A Theory of Negligence*, 1 J. LEGAL STUD. 29, 33 (1972) ("One misses any reference to accident avoidance by the victim. If the accident could be prevented by the installation of safety equipment or the curtailment or discontinuance of the underlying activity by the victim at lower cost than any measure taken by the injurer would

does not hold potential plaintiffs liable for failing to anticipate the potential negligence of others.[99] But we are practically able to control the costs of accidents to ourselves via our knowledge of others.[100] Where this knowledge breaks down, an important element of the tort system fails to operate. We offer five examples to justify this point.

First, products liability law has made manufacturers liable not only for intended use of their products but for foreseeable misuse of their products.[101] The idea here is to not let manufacturers dodge liability by claiming they intended for the products being used in one way even though they knew that people will commonly use it in a different way. Tort loss thus holds manufacturers responsible for making judgments about how ordinary people will use their products and taking reasonable precautions to prevent harmful misuses.[102] This standard relies on the ability of manufacturers to make those judgments about how ordinary people will interact with their products.

Second, vicarious liability for the acts of agents also incorporates a sense of whether the tort was foreseeable. Respondeat superior claims which impose the strictest liability still require the employee to have been acting within the scope of their employment and the doctrine is justified, in part, by the view that firms can choose who to hire, how to supervise them, and what tasks to have them do such that the risk to others is limited.[103] Agency law also makes direct liability

---

involve, it would be uneconomical to adopt a rule of liability that placed the burden of accident prevention on the injurer.").

[99] *See* Andrew S. Gold & Henry E. Smith, *Sizing Up Private Law*, 70 U. TOR. L.J. 489, 503 (2020) (using the example of engine spark cases to discuss how this rule generates local simplicity in tort law but frustrates some modern commentators concerned with efficiency); *see also* Mark F. Grady, *Common Law Control of Strategic Behavior: Railroad Sparks and the Farmer*, 17 J. LEGAL STUD. 15, 18-19 (1988) (providing analysis of common law rules asking plaintiffs to take precautions in response to the negligence of others); Posner, *supra* note 98, at 60 (arguing that the treatment of engine sparks was efficient

[100] When we see a driver recklessly swerving through traffic on the highway we tend to slow down and let them pass rather than rely on the tort system for our protection.

[101] RESTATEMENT (THIRD), *supra* note 24, § 2 R.N. ("When a product is put to an unforeseeable use and the plaintiff claims that the product should have been designed to avoid injury when put to such a use, the courts agree that liability will not attach. There is widespread understanding that misuse in this context goes to the basic duty issue.").

[102] *Id.* § 2 cmt. m ("Subsections (b) and (c) impose liability only when the product is put to uses that it is reasonable to expect a seller or distributor to foresee. Product sellers and distributors are not required to foresee and take precautions against every conceivable mode of use and abuse to which their products might be put. Increasing the costs of designing and marketing products in order to avoid the consequences of unreasonable modes of use is not required.").

[103] *See* RESTATEMENT (THIRD) OF AGENCY § 2.04 cmt. b (AM. L. INST. 2006) (discussing rationales for the doctrine of *respondeat superior*). In particular, the Restatement points out that the firm or principal has the ability to both limit employee discretion and structure their work so as to limit the risk to others. *Id.* ("A firm or organization that employs individuals usually structures their work to limit the scope of discretion and individual action, thus limiting the occasions when unreasonable decisions are likely to be made. . . . Respondeat superior creates an incentive for

available where the tort is in some way specifically foreseeable by the employer in the sense that they had reason to know the employee would engage in the conduct in question.[104]

Third, claims for negligent supervision of children typically require a plaintiff to prove (1) that the parent was aware of the potential misdeeds and (2) that they had the ability to control the child.[105] This first prong again reflects a human assumption. We will not hold parents liable for errors committed by their children that are truly unpredictable because we believe parents are entitled to rely on the social expectation that children do not ordinarily commit torts. But once parents are on notice about potential issues, then responsibility shifts for them to exercise control where they are capable. In essence, once the errors become foreseeable, a duty to make reasonable efforts to control emerges.

Fourth, the same is true for questions about controlling animals. Animals, of course, are not human and thus do not harm others in a way that a human would. Nevertheless, a similar foreseeability standard is also implicitly embedded in the rules governing harms due to animals that again, we think, fits with social expectations of how animals tend to behave. We can see this dynamic in the trio of rules governing domestic and wild animals. First, with respect to animals who are not known to be dangerous, the law imposes only a negligence duty to prevent harm by the animal.[106] Second, if a domestic animal is known to be abnormally dangerous, liability becomes strict albeit only for the known dangerous propensity.[107] Third and finally, liability for wild animals is strict for

---

principals to choose employees and structure work within the organization so as to reduce the incidence of tortious conduct."); *see also* Diamantis, *Reasonable AI*, *supra* note 2, at 585 ("The final element of respondeat superior-fault-is the trickiest to adapt to AI. With human employees, fault poses little mystery. We are familiar with various ways that humans can fall short, many of which have been codified in law for centuries." (footnote omitted)).

[104] *Id.* § 7.05(1) ("A principal who conducts an activity through an agent is subject to liability for harm to a third party caused by the agent's conduct if the harm was caused by the principal's negligence in selecting, training, retaining, supervising, or otherwise controlling the agent."). Once more the Restatement ties liability to the foreseeability of the harm among other considerations. *Id.* § 7.05 cmt. d ("Conduct that results in harm to a third person is not negligent or reckless unless there is a foreseeable likelihood that harm will result from the conduct. A determination that a person acted negligently also requires consideration of the foreseeable severity of harm to a third person. When a principal conducts an activity through another person, the nature of the task to be performed and the conduct required for performance are relevant to whether the principal acted negligently, either in selecting the actor or in instructing, supervising, or otherwise controlling the actor." (citation omitted)).

[105] RESTATEMENT (SECOND), *supra* note 69, § 316. The Restatement notes that this rule relaxes the formerly strict liability for the father as head of household for the actions of members of his household and instead asks both parents to exercise responsibility for behavior they have the ability to control. *Id.* § 316 cmt. a.

[106] *Id.* § 518. And, of course, to not intentionally use the animal to do harm. *Id.*

[107] *Id.* § 509.

the characteristics of wild animals of this class.[108] Here again we can see a structure of liability based on common knowledge. Domestic animals are typically safe, and their owners can assume as much, exercising ordinary care to prevent any harm. Owners who are on notice of the danger of their pets then are strictly liable for these dangers. Keepers of wild animals are subject to the strictest liability on the grounds that they have created an *abnormal* and thus unexpected danger to others.[109] But in keeping with foreseeability, strict liability only attaches to accidents caused by said abnormal danger (though the possessor is *required* to know of the nature of the danger).[110]

Like an artificial intelligence system, wild animals harm others in ways that are unpredictable. Even where the law does not forbid possession of said animals, owners act at their own peril in keeping said animals even when they have a reasonable belief they have tamed them and take utmost precautions to contain them.[111] Faced with the choice of letting the losses lie with the victims or with the owners, liability is assigned to the owner as a way to deter inviting this type of randomness into the community. Though children and dogs also are prone to a certain type of randomness in their harmful actions, the expectation is that most children and most dogs are not dangerous and owners are permitted to rely on this commonly shared assumption. Moreover, the attractiveness of this rule is buttressed by both the social benefit of children and domestic pets as well as well-developed expectations of how to avoid harms due to the clumsiness of children or the unexpected actions of pets.[112]

Fifth, contributory and comparative fault also reflect a similar balancing between actors. Particularly under a regime of contributory negligence or modified comparative fault, defendants can be assured that their lapses will not give rise to liability if plaintiff is mostly at fault for the accident. When considering which precautions to take, defendants (and, of course, plaintiffs) can assume that their counterparties also have reason to take care. Though we doubt that individuals actually engage in an extensive consideration of how much negligence they can get away with and still avoid liability on the grounds that they can assert comparative or contributory negligence defenses, we again think the concern for comparative fault reflects an interest in distributing fault fairly

---

[108] *Id.* § 507.

[109] *Id.* § 507 cmt. e ("The rule stated in this Section is based upon the fact that by keeping a wild animal of a class that has dangerous propensities, its possessor has created a danger not normal to the locality in question.").

[110] *Id.* § 507 cmt. c ("[The possessor] may reasonably believe that it has been so tamed as to have lost all of these [dangerous] propensities; nonetheless he takes the risk that at any moment the animal may revert to and exhibit them. Therefore, a keeper of an elephant that for years has shown none of the dangerous traits common to elephants as a class is liable under the rule stated in this Section if the elephant suddenly and unexpectedly exhibits these traits.").

[111] *Id.* § 507 ill. 1.

[112] Think, for example, of the norm taught to children that one should ask the owner of a dog whether the dog will welcome being approached and pet.

between human actors who are certain to, on occasion, fall short of reasonable care in human ways. To the extent that machine carelessness is unlike human carelessness, this balance is upset.

Finally, we note that outside of these examples specific negligence law also embeds this type of humanistic assumption in the fundamental doctrines of negligence law itself. Questions of duty, [113] breach,[114] and proximate causation[115] all involve considerations of foreseeability. One way to justify those requirements is to say that they require individuals when structuring their affairs to think carefully about how they might harm others and to take reasonable efforts to prevent said harms. But individuals do not act in a vacuum and whether or not a particular course of conduct creates a risk to others depends, in part, on the actions of other actors. Thus, the *Restatement (Second)*, when discussing the standard for breach, states that an actor is "required to know . . . the qualities and habits of human beings and animals and the qualities, characteristics, and capacities of things and forces in so far as they are matters of common knowledge at the time and in the community . . . ."[116]

Negligence requires us to use our knowledge of ourselves and the ways in which we are fallible to control our behavior and prevent it from imposing unjustifiable risks on others. It also asks us to consider our common knowledge of the qualities of other people, animals, and societal dynamics when choosing what care to take. The law at a fundamental level asks us to make a judgment about the expected ways in which humans err and, in particular, the ways in which we, personally, err.

Reflecting that judgment, tort law will also not let experts off the hook on the grounds that a person without their expertise would have also inevitably caused the same accident.[117] Thus, a surgeon cannot point to the fact that an ordinary person would also fail to perform a surgery adequately if pressed into action to avoid a claim of malpractice. Where a more tailored expectation on behalf of the individual about how they might err and harm another is available, the law

---

[113] RESTATEMENT (SECOND), *supra* note 69, § 281 cmt. b ("In order for the actor to be negligent with respect to the other, his conduct must create a recognizable risk of harm to the other individually, or to a class of persons—as, for example, all persons within a given area of danger—of which the other is a member.").

[114] *Id.* §§ 283, 289-90 (discussing the knowledge the reasonable man is assumed to possess when determining the scope of his or her duty).

[115] *See, e.g.*, Garcia v. Col. Cab. Co., 538 P.3d 328, 333 (Colo. 2023) (linking foreseeability and proximate cause); *see also* RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 29 (AM. L. INST. 2010) ("Courts have increasingly moved toward adopting a foreseeability test for scope of liability in negligence cases. Currently, virtually all jurisdictions employ a foreseeability (or risk) standard for some range of scope-of-liability issues in negligence cases.").

[116] RESTATEMENT (SECOND), *supra* note 69, § 290.

[117] *Id.* § 289(b) (incorporating into the standard of care "such superior attention, perception, memory, knowledge, intelligence, and judgment as the actor himself has").

is willing to incorporate that understanding and ask that we understand ourselves.

One might object that the foregoing analysis mushes together questions of duty, breach, and proximate cause. The problem is arguably exacerbated by combining a discussion of *respondeat superior*, a vicarious theory of liability with direct theories of liability against principals. It's a fair point. In some cases, the foreseeability assumptions we identify are tied to duty. For example, the negligent supervision standard sounds in duty and breach while the limitation on liability for injuries by wild animals sounds in proximate cause (as strict liability sets an unlimited duty of care). With respect to the first point, we accept that we have moved seamlessly between the elements without acknowledging the movement explicitly. Our aim, however, is not a sleight of hand but to point out that foreseeability notions are present in all of these doctrines from the strictest standard of liability to mere negligence even if it rears its head in different parts of the analysis and takes a different form in different elements.[118] Without committing to a particular philosophical foundation for tort law, our suggestion is that the appearance of foreseeability and its incorporation of known social norms grounds the fairness of these tort doctrines for both defendants and plaintiffs. Our response to the second point about vicarious liability is much the same. While *respondeat superior* as a theory of vicarious liability is fundamentally different from the duty imposed by a direct theory of liability, our point is that the rationales for the doctrine still sound in a light form of foreseeability that justifies the doctrine's fairness in terms of the employer's ability to select, control, and monitor employees.[119] To be sure, the strictness of this form of vicarious liability will, at times, reach beyond foreseeability. And some commentators might simply justify it on a cost of doing business rationale that resembles insurance.[120] Nevertheless, we believe the foreseeability/fairness account best captures the difficulty artificial employees pose for the doctrine.

---

[118] *Cf.* Benjamin C. Zipursky, *Foreseeability in Breach, Duty and Proximate Cause*, 44 WAKE FOREST L. REV. 1247, 1247-49, 1271-75 (2009) (critiquing the *Restatement (Third)*'s attempt to attempt to limit consideration of foreseeability in duty and replace foreseeability in proximate cause, though providing some support for its attempt to reformulate proximate cause); *see supra* notes 113-116 and accompanying text.

[119] In this sense, there may be a loose analogy to the case of wild animals in which the unlimited duty imposed by strict liability is perhaps justified on a similar choice of the defendant to enter the arena and the fairness of asking them to carefully select, control, and monitor.

[120] *See* Taber v. Maine, 67 F.3d 1029, 1036-37 (2d Cir. 1995) (Calabresi, J.) (summarizing the modern view that *respondeat superior* is concerned with spreading the costs of business injuries and thus concerned with whether the employee's conduct is so unusual that it is unfair to consider it a cost of the employer's business). Though even this treatment by Judge Calabresi includes consideration of the relationship between the employee and the employer and what the employer can foresee as a cost of doing business. *Id.*

A second objection could be made that foreseeability is a malleable term that often folds in or expresses our considered judgements about what is fair.[121] If so then the argument is somewhat circular since foreseeability is just a statement of fairness. That is, decisions that this or that harm is foreseeable are just statements about whether it is good policy or fair to recognize liability in a given case.[122] We agree to some extent but note that even if we accept this characterization, we can still argue for the fairness of the above doctrines and what they expect defendants to do and plaintiffs to bear in terms of social knowledge and assumptions about how others behave. Confronted with truly unpredictable agents, fairness/foreseeability is undercut in each of these cases rendering a single doctrinal box an imperfect fit.

For example, consider an artificial employee "hired" from an AI developer who errs in a small number of cases in an unusual and perhaps catastrophic way. The novelty of the technology makes it hard for the employer to monitor and control the employee. How does one do so if the artificial employee fails even if given the same instructions as a human employee. Similarly, given the low rate of failure it seems unreasonable to straightforwardly claim that the employer had reason to know when selecting the artificial employee that it was likely to err (it wasn't). Application of *respondeat superior* in such a case might be appropriate but seems to us to resemble insurance unmoored from the rationales described above.[123]

In sum, when it comes to controlling the actions of others or our own actions, the law requires us to make common sense judgments about the world and the ways in which our actions can potentially harm others. This inquiry relies explicitly on judgments about how humans or animals typically behave both at the level of individual, specialized doctrines and at the level of our general duty of care. Because most children and most household pets are not dangerous, the law permits us to rely on that assumption. Once we know otherwise, the law shifts and requires us to control both animals and children where possible. Similarly, in asking how we should structure our own affairs to avoid imposing unjustifiable risk on others, the law asks us to consider both our own capabilities and those of the reasonable person that we acquire based on living in society and

---

[121] John Fabian Witt & Margan Savige, *Foreseeability Conventions*, 44 CARDOZO L. REV. 1076, 1088-94 (2023) (summarizing the skeptical view of foreseeability as indeterminate and malleable); *see* Tarasoff v. Regents of Univ. of Cal., 551 P.2d 334, 342-343 (Cal. 1976) (stating the rule for duty as a balance of factors of which foreseeability is the most important); *see also* Zipursky, *supra* note 118, at 1258-60 n.47 (collecting cases on the role of foreseeability in duty from state supreme courts, many of which also call for a consideration of public policy or fairness).

[122] This can be true whether the foreseeability in question is being deployed in the duty, breach, or proximate cause prong.

[123] Indeed, this hypothetical scenario seems to fail Diamantis's control test. *See supra* notes 71-73 and accompanying text.

seeing other people around us and to take reasonable precautions based on that understanding.

The difficulty, however, is that when those expectations are absent—as in the case of many artificial intelligence models—then the comprehensibility of these doctrines, we think, comes somewhat undone. That problem is not necessarily unfixable, but it poses a problem for straightforward application of tort law. It's tempting then to apply the solution that we find with wild animals which are both random and dangerous. The trouble is that many mundane uses of artificial intelligence are not dangerous in the way that keeping a puma in one's house might be. Further, as such models spread, the risk posed by artificial intelligence is not abnormal and unusual, it is abnormal and a part of everyday life.[124] Perhaps more importantly, we believe that artificial intelligence has the potential to improve human life substantially.[125] It thus seems inappropriate to treat AI akin to the keeping of wild animals or as an abnormally dangerous activity.[126] The result is ambiguity about where AI fits within the scheme of tort law.

B. AI Failure and the "Foreseeability Gap"

To be sure, the law does appreciate that some errors are truly random. If driver *A* has a stroke and crashes into driver *B* then losses are likely to lie where they fall.[127] But as the preceding Section argues, these gaps are unexpected but not unusual. They are beyond the foresight of both the reasonable injurer and the reasonable victim. Accordingly, we can perhaps imagine that the injurer and

---

[124] *See* source cited *supra* note 29 and accompanying text.

[125] *See, e.g.*, ABBOTT, *supra* note 2, at 55-56 (summarizing sources who believe that widespread adoption of autonomous vehicles could eliminate 90% of deaths due to automobile accidents). This is to say nothing of more speculative uses such as AI drug discovery. *See* Anthony King, *Four Ways to Power-Up AI for Drug Discovery*, NATURE (Feb. 27, 2025), https://www.nature.com/articles/d41586-025-00602-5.

[126] *See* RESTATEMENT (SECOND), *supra* note 69, § 519 (providing for strict liability for abnormally dangerous activities). As with wild animals, the *Restatement (Second)* justifies this stance by stressing out the abnormality of the danger. *Id.* § 519 cmt. d ("It is founded upon a policy of the law that imposes upon anyone who for his own purposes creates an abnormal risk of harm to his neighbors, the responsibility of relieving against that harm when it does in fact occur."). This rule and the rule for wild animals strikes us as a thumb on the scale against behavior of questionable social value and indeed this consideration is reflected in the definition of an abnormally dangerous activity. *Id.* § 520 (considering, among other factors, the "inappropriateness of the activity to the place where [the activity] is carried on" and the "extent to which [the activity's] value to the community is outweighed by its dangerous attributes"). Focused on autonomous vehicles, Abbott thus argues that strict liability treats robot drivers worse than human drivers and therefore deters adoption of a socially valuable technology. ABBOTT, *supra* note 2, at 57-58.

[127] That the losses are born by auto insurance in this particular case would be one benefit of having such a system of mandatory insurance even if it is not the primary motivation.

injured could agree that neither will be responsible for the accident and the losses will lie where they may.[128] Moreover, because these cases are well understood, insurance can fill the gap.

The promise of artificial intelligence is that it will introduce a set of "actors" who often err in odd ways into ordinary life and commerce. This dynamic introduces what we will call the "foreseeability gap" into tort law. A rigorous application of foreseeability principles may result in a large number of accidents that go uncompensated if factfinders conclude that such failures are truly random.

To be sure, scholars have argued otherwise. As noted, Anat Lior notes that the applicability of foreseeability to AI system depends on the level of generality at which the concept is applied.[129] While it is difficult to predict what types of errors will emerge from an AI system it is predictable that they will err.[130] If courts choose to adopt a view of foreseeability that operates at a high level of generality then AI manufacturers will be asked to operate as *de facto* insurers. Such a system, however, would be subject to Vladeck's critique that manufacturer liability does not properly account for the reality of AI supply chains.[131]

The further trouble is that if we accept that foreseeability is evaluated at the level of deployment of AI the result is a system of liability that seems out of step with both fairness and economic good sense. With respect to the former, strict liability in this space seems out of step with other areas in which it is deployed. Whatever one's quibbles with AI systems, their use is not questionable in the way that inherently dangerous activities or the keeping of wild animals are. Turning to the latter, the potential productivity and safety gains of thoughtfully deployed artificial intelligence systems seem sufficiently great that a system of liability that, in essence, deters the development and deployment of AI seems short-sighted to us.

On the other hand, a low level of generality for foreseeability that opens a large foreseeability gap also seems problematic. Even if we admit that AI will decrease the number of accidents on net—a claim that seems to depend on the system at hand—a large foreseeability gap would seem to raise some of the problems that bedeviled Industrial Revolution tort law that left too many victims uncompensated in the name of *laissez faire* ideology. Perhaps that's the best we can do, or we need to choose between undesirable options of broad or narrow liability.

---

[128] Indeed, contract law's default rules will release a party from performance of a contractual duty if "after a contract is made, a party's performance is made impracticable without his fault by the occurrence of an event the non-occurrence of which was a basic assumption on which the contract was made." RESTATEMENT (SECOND) OF CONTRACTS § 261 (AM. L. INST. 1981).

[129] Lior, *supra* note 11, at *3-*5.

[130] *Id.* at *4.

[131] *See supra* Section I.B.

Our contention is that we can do better. By adopting a mid-level solution, we can help bridge the gap. Our solution will not eliminate all of the problems of foreseeability just as traditional tort law does not. But by focusing on whether an error arose from a known *reason* (and considering the culpability of the user and their *notice*) we believe we can close the foreseeability gap at least somewhat.

* * *

To summarize the argument to this point, we have argued that tort law embeds questions of foreseeability for both fairness and deterrence reasons. Moreover, the way in which foreseeability is deployed—even where the law imposes strict liability—subtly depends on social understandings about the way in which the world works. Artificial intelligence systems pose a problem for this paradigm because even if they err less often than the reasonable person, they err in strange ways. Because social understandings about these errors have not developed and because the development of AI systems seems pro-social in general, we want to seek out a solution that adapts tort law in a way that can compensate victims and does not take an overly harsh approach to the industry. Our solution is to look at the *ways* and *mechanisms* that drive AI failure. The next Part thus turns to the question of how and why AI fails.

III. HOW AND WHY AI FAILS

This Part focuses on the nature of AI failure. We argue that by understanding AI as fundamentally an engine for predicting the "best" response to some set of inputs, we can see that there are several well-known *reasons* why AI systems fail. We present this analysis in three Sections. First, we discuss a handful of known reasons that can cause AI systems to fail. We note that despite enumerating these foreseeable types of errors, they were almost all at one point unforeseeable. Second, we discuss some of these failures in real-world scenarios to show that the same type of error can produce different liability outcomes. Although we focus on specific use cases of artificial intelligence, the ideas generalize to other modalities and types of AI. Third, we reference the work made by the academic community to close this foreseeability gap.

A. How AI Errs

i. Terminology and Notations

We largely consider the artificial intelligence and machine learning problem in this paper as one of inputs and outputs. A model f_θ takes inputs X from the

surrounding environment and produces outputs $\hat{y}$ (the hat is used to denote that this value is an estimate or prediction produced by the model). Models in machine learning can take many forms such as random forest, decision trees, and, most notably now, deep neural networks. For the purposes of this work, we are agnostic to the inner workings of the model f_θ and how it is trained with the caveat that perversions in the training apparatus (such as using biased data or a bad loss function) can cause negative downstream effects requiring liability analysis.[132]

We briefly define some notation. A model f_θ operates in an environment or dataset D and has learned internal parameters θ. The model takes as input $X_i \in X$ from the environment and produces outputs $\hat{y} \in Y$. A given input X_i may or may not have a corresponding ground truth value y_i (note the lack of "hat" on the ground truth value). Practitioners measure the performance of the model using an objective function (sometimes called a *loss function* in the ML literature) and update the internal parameters of the model, thus improving the performance of the model. In the view of machine learning transforming inputs to outputs, we can generalize in notational form $f_\theta(X) = y$ or $f_\theta: X \to Y$. During training, we take inputs sampled from D_{train} that should be distinct from the samples during test time, D_{test}. This assumes an idealized world where D is known and non-changing. However, AI developers often construct a dataset and create a world model D that deviates from the true world D^*. Furthermore, the world is not a static place and evolves over time. We use the subscript $t \in [0, T)$ to denote a version or snapshot of both the constructed and true worlds D_t and D_t^*.

We begin with a brief discussion on how the training apparatus can fail leading to disastrous downstream effects during deployment. We follow with problems introducing AI systems into an "open-world" that evolves constantly and may deviate from the limited environment introduced during training. Lastly, we focus on errors that arise during deployment even in the most conducive environments.

---

[132] There are generally three main learning paradigms in modern machine learning. First, supervised learning trains on a set of labeled input data X and output data y. During training, model M takes the training data X to produce predictions $\hat{y}$. Some learning mechanism adjusts internal model parameters θ based on the measured difference between y and $\hat{y}$. Second, reinforcement learning creates a world model and an objective function on the current state of the world. The machine learning model changes the state of the world, measures the new value of the objective function, and accordingly updates internal model weights, θ. Third, unsupervised models train on a set of unlabeled data X, often by trying to mimic the data, and measure the difference between the real and created data to update internal model weights, θ. There are other learning paradigms that often mix components of each of these three such as active learning, federated learning, semi-supervised learning, etc. For the purposes of our paper, we are agnostic to the learning paradigm and treat machine learning and AI systems collectively as tools that transform inputs X into outputs $\hat{y}$.

ii. Learned Model Bias

Modern machine learning techniques require a large amount of data. Some estimate that GPT-4 required over thirteen trillion tokens[133] of input data during training.[134] With such large amounts of data during training, the adage *garbage in garbage out* remains as strong as ever.[135] Furthermore, the sheer size of data means that humans cannot adequately monitor the inputs to these models.[136] The models are biased towards the distribution of data received during training. If the data is biased, the models will be as well. Melanoma is a much-studied deadly skin cancer. Researchers in artificial intelligence have devoted a significant amount of time to the detection of melanoma in images of skin lesions.[137] People with lighter skin tones tend to suffer from higher incidences of melanoma than those with darker complexions. However, melanoma can develop in anyone. Because of the discrepancy in likely cases, and the racial demographics of countries most likely to produce large-scale medical datasets, the preponderance of available samples for training both positive and negative are of lighter skinned individuals. Consequently, the models themselves tend to have a built-in bias and have significantly higher classification accuracies on a narrow subset of skin tones.[138] In this example, biases in the training data lead to biases in the model outputs.

---

[133] Large language models first transform text into tokens. Tokens represent sub-words. So "surprisingly" might be split into three tokens: "surpris" "ing", and "ly". Here, "surpris" represents the root word, "ing" indicates a present participle or gerund form, and "ly" indicates an adverb.

[134] Stephen M. Walker II, *Everything We Know About GPT-4*, KLU (September 1, 2025), https://klu.ai/blog/gpt-4-llm.

[135] The topic of model bias, particularly in criminal law, has received a large amount of attention in the legal literature as well. *See, e.g.*, CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019); Andrés Páez, *Negligent Algorithmic Discrimination*, 84 L. & CONTEMP. PROBS. 19 (2021; Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109 (2017). Because our focus in this Part is on the computer science of bias, we focus here on that literature but want to acknowledge the wide legal literature on this feature of artificial prediction.

[136] Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 610 (2021), https://dl.acm.org/doi/10.1145/3442188.3445922

[137] SangHyuk Kim et al., *Melanoma Detection With Uncertainty Quantification,* 2025 IEEE INTERNATIONAL SYMPOSIUM ON BIOMEDICAL IMAGING (2025), https://arxiv.org/abs/2411.10322.

[138] Laura N. Montoya et al., *Towards Fairness in AI for Melanoma Detection: Systemic Review and Recommendations*, FUTURE OF INFORMATION AND COMMUNICATIONS CONFERENCE (2025), https://arxiv.org/abs/2411.12846.

These problems exist in a wide-ranging set of fields from criminal justice,[139] healthcare,[140] and insurance outcomes.[141] Data from a historically biased world will produce models with the same learned biases. This produces significant challenges for artificial intelligence companies which must comply with state and federal civil rights legislation.[142] Judges that use COMPAS, for example, want to ensure that the recidivism prediction uses non-racial reasoning.[143] Furthermore, the models should not decide based on variables that are racially coded. For example, recidivism prediction that uses neighborhood or ZIP code as a proxy for race in highly segregated cities across the country would be a significant issue.[144]

Explainability is an active area of research that asks models to produce a value (prediction) as well as reasoning for that prediction.[145] Explainability can reduce the risk of model biases if there is a human-in-the-loop evaluating the correctness of the outputs of the model (e.g., a judge who uses COMPAS as a guide for sentencing as opposed to an oracle). However, explainability is an active area of research because it is *hard*. The trend of artificial intelligence has been deeper neural networks with an ever-increasing number of parameters. As these networks become more complex, explainability becomes even harder.[146] Models deployed in the real world are also unlikely to always have human supervision. Thus, even with perfect explainability a biased model will produce records for audit after the infraction has occurred. Although better than nothing, it still has the very negative consequence that bias is perpetuated in the short term.

---

[139] Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[140] Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447, 447 (2019).

[141] Fei Huang, *How Insurers Can Mitigate the Discrimination Risks Posed by AI*, I BY IMD, (July 18, 2022), https://www.imd.org/ibyimd/innovation/how-insurers-can-mitigate-the-discrimination-risks-posed-by-ai/.

[142] One difficulty that is well-known in this space is that there are multiple plausible metrics of racial equality and it is typically not possible for an algorithm to meet all at once. Mayson, *supra* note 135, at 2233-38 (discussing the problem including its relevance to the dispute between the makers of COMPAS and ProPublica).

[143] Angwin et al., *supra* note 139.

[144] The exact variables used by COMPAS are not known as the model is proprietary. See Robin A. Smith, *Opening the Lid on Criminal Sentencing Software*, DUKE TODAY (July 19, 2017), https://today.duke.edu/2017/07/opening-lid-criminal-sentencing-software.

[145] Nadia Burkart & Marco F. Huber, *A Survey on the Explainability of Supervised Machine Learning,* 70 J. ARTIFICIAL INTELLIGENCE RES. 245, 245-248 (2021) (providing an overview of the topic)

[146] Fenglei Fan et al., *On Interpretability of Artificial Neural Networks: A Survey*, IEEE TRANSACTIONS ON RADIATION AND PLASMA MEDICAL SCIENCES (2021).

iii. Reward Hacking

Developers (often) train their machine learning models using objective or loss functions.[147] Simply, the developers measure the difference between the expected and realized output from the model and update the internal parameters accordingly. In this way, the model gradually improves its performance during training. Occasionally, the model can learn shortcuts or abnormal behavior that optimize the loss function in unexpected ways. Consider the example previously discussed from OpenAI of the boat-racing video game.[148] In the game, CoastRunners, the objective is to finish the game quickly while simultaneously hitting objects along the route. As the objective function, the developers simply used the score output from the video game. Unfortunately, the score from the video game does not reward proximity to the finish line. Thus, the AI agent "learned" that optimal behavior was to continually hit objects without progressing towards the finish line.

The developers discovered this humorous example in the laboratory after watching the agent converge to this optimal behavior in a simple game setting. However, consider a more complex training set up where a company seeks to produce an autonomous vacuum. The developers use a reinforcement learning model with an objective function that rewards the absence of dust and debris from the camera mounted on top of the vacuum. The developers train the robot on relatively de-cluttered houses and apartments. Occasionally, the vacuum wedges itself into dark nooks and crannies and abruptly stops. The developers pay no attention to these iterations and restart the vacuum in a different part of the house. What the developers fail to realize is that the vacuum is learning that darkness obscures the camera and thus removes any sight of dust from its camera (producing an optimal output from the objective function). When the developers put their product to market, customers (perhaps living in more cluttered environments) complain that the robot simply finds the nearest pile of clothes or blankets and hides in it.

iv. Out-of-Distribution Inputs

Machine learning models, especially deep neural networks, perform incredibly well on withheld testing data on in-distribution samples. For example, a model trained to classify breeds of dogs will achieve high accuracy when provided with new images of known (trained-on) dog breeds. The success of these

---

[147] There are some learning paradigms that do not require objective or loss functions either through the use of randomization or simpler "splitting" criteria such as decision trees. However, the trend in machine learning has (for several decades) been towards loss functions to update model parameters. This is the major learning mechanism employed for deep neural networks and is thus extremely relevant for AVs and LLMs.

[148] Clark & Amodei, *supra* note 12.

models is most notably seen on the large-scale benchmarks produced by the academic community. The ImageNet Large Scale Visual Recognition Challenge contains 1000 object classes with between 732-1300 samples per class in the training dataset for a whopping 1,281,167 human labeled training images.[149] State-of-the-art solutions such as CoCa achieve 91% top-1 classification accuracy on the ImageNet testing dataset. [150]

It is not surprising that a model trained to differentiate cars from airplanes might fail when provided with a battleship. However, the models do not just fail to classify the input as "battleship"[151]—the model will, with high confidence, predict the battleship as either a car or an airplane![152] As models have increased in the number of parameters, the calibration has often worsened.[153] By calibration, we mean that the model is wrongly (but often overly) confident in its predictions. A properly calibrated model would provide a confidence score of *X*% on samples that, when aggregated, are correct around *X*% of the time. The problem is often exacerbated on these out-of-distribution inputs.[154]

There has been significant research into detecting out-of-distribution inputs before, during, and after inference.[155] However, the area of research remains active with no set solution outperforming all others. As such, we cannot expect commercial AI systems to adequately mitigate this problem. AI models may and often will, when presented with out-of-distribution samples, fail in difficult to detect ways.

---

[149] *See* Jia Deng et al., *ImageNet: A Large-Scale Hierarchical Image Database,* IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITIONS (2009); Olga Russakovsky et al., *ImageNet Large Scale Visual Recognition Challenge*, 115 INTL. J. COMP. VISION 211 (December 2015).

[150] Jiahui Yu et al., *Coca: Contrastive Captioners Are Image-Text Foundation Models* ARXIV (June 14, 2022), https://arxiv.org/pdf/2205.01917

[151] Which, again, is not surprising, since the model outputs are constrained by what it has seen.

[152] A brief note about confidence: a machine learning model takes inputs *X* to outputs *y*. In classification tasks, *y* is often a vector where each element of the vector corresponds to a different class within the classification problem. So, for example, the first element of the output vector might correspond to the likelihood of a car, the second to airplane, and so on. A model is said to have high confidence if the most likely class (the element with the highest value in the vector) is significantly greater than the other classes. A very simple method for measuring confidence is the softmax function.

[153] Chuan Guo et al., *On Calibration of Modern Neural Networks,* PROCEEDINGS OF THE 34TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING (2017), https://arxiv.org/pdf/1706.04599

[154] Christian Tomani et al., *Post-hoc Uncertainty Calibration for Domain Drift Scenarios*, IEEE/CVF CONFERENCE ON COMPUTER VISION (2021), https://arxiv.org/abs/2012.10988.

[155] *See* Jingkang Yang et al., *Generalized Out-of-Distribution Detection: A Survey,* 132 INTL. J. COMP. VISION 5635 (2024).

v. Concept Drift

We discussed previously that models are trained on a constructed version of the world (sometimes a curated dataset) that represents a snapshot of time D_0. However, the world evolves constantly and with it the distribution of data. Let's consider classes C representing different types of internet traffic ranging from benign to known attacks such as distributed denial-of-service or Heartbleed.[156] A cybersecurity-focused machine learning model might look at network traffic between a host network and the outside world and predict if the incoming traffic is benign or malicious. Malicious traffic can then be routinely discarded before the payload degrades the performance or compromises the integrity of the host network. The model will rely on labeled data from a dataset D_0 taken from a given year. However, the internet is constantly evolving, and with it the appearance of each class. The benign class, for example, would typically contain unencrypted HyperText Transfer Protocol (HTTP) connections before 2015. By 2017, most benign traffic follows the HyperText Transfer Protocol Secure (HTTPS) which provides a level of encryption.[157] A model trained to monitor traffic before 2015 would quickly flounder by 2017.[158]

We refer to this world evolution as concept drift, and it poses a large problem for machine learning models. As the data distribution changes, the guarantees and expected performance of the models will degrade. This phenomenon has been discussed in the academic literature for quite some time in numerous fields such as medicine, cybersecurity, and natural language.[159] Of particular interest are machine learning products that are purchased without an ongoing subscription. These models will almost inevitably degrade in performance eventually as the world (and thus the inputs to the model) evolves. Without a subscription service that provides model updates, the model will remain stagnant. If the manufacturer of such a model is found liable for an inevitable eventual failure, it almost necessitates subscription pricing for every commercially available machine learning tool.

---

[156] For examples of data sets, see Iman Sharafaldin et al., *Towards Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization*, INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS SECURITY AND PRIVACY (2018); and Nour Moustafa et al., *UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set)*, MILITARY COMMUNICATIONS AND INFORMATION SYSTEMS CONFERENCE (2015).

[157] *See Let's Encrypt Stats*, LET'S ENCRYPT, https://letsencrypt.org/stats/ (Accessed: January 20, 2026).

[158] Brian Matejek et al., *SAFE-NID: Self-Attention With Normalizing Flow Encodings for Network Intrusion Detection*, TRANSACTIONS ON MACHINE LEARNING RESEARCH (2025).

[159] Pang Wei Koh et al., *WILDS: A Benchmark of In-The-Wild Distribution Shifts*, PROCEEDINGS OF THE 38TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING (2021).

vi. Continual Learning and Catastrophic Forgetting

An obvious question when presented with the problems of concept drift is "Why don't artificial intelligence producers continually update their model weights in the presence of an ever-changing world?" The answer is that continual learning is an incredibly difficult problem that can induce what researchers call "catastrophic forgetting".[160] During catastrophic forgetting, a model "forgets" the previous training distribution and struggles to correctly classify examples from that distribution. Ideally, the in-distribution samples of the model will be an ever-increasing set but in practice the model's distribution becomes warped. There is a significant amount of research on how to prevent catastrophic forgetting, but this is far from a solved problem.[161]

Catastrophic forgetting is particularly challenging when a model trained on a given task is updated to perform a novel task (in addition to the old one).[162] This will become an increasing problem as industry favors general models that can accomplish many tasks as opposed to specialized models that perform a singular task. What practitioners often find is that the re-training updates model weights that were critical for performing the original task. Thus, the model will learn to perform the new task while forgetting how to perform the original task.

vii. Human Alignment Failures

The major artificial intelligence companies today use a technique called reinforcement learning from human feedback (RLHF) to finetune their language models. Partly, this learning forces the models to be helpful. When queried, GPT tries to provide an answer rather than raw information. This is learned behavior after a large number of human evaluators have interacted with the base model and produced positive feedback on "helpful" answers. Another critical component of RLHF is aligning models with human (often liberal, democratic) values.[163] Perhaps more accurately, Western companies align their models with liberal democratic values. Thus, a model will rather not produce racist and/or authoritarian outputs. DeepSeek, a Chinese company, aligns their models with

---

[160] ZHIYAN CHEN & BING LIU, LIFELONG MACHINE LEARNING 55-57 (2018) (proving an overview of the problem).

[161] Wickramasinghe *et al.*, *Continual Learning: A Review of Techniques, Challenges, and Future Directions*, 5 IEEE TRANSACTIONS ARTIFICIAL INTELLIGENCE 2526 (2023) ("In modern AI, [Continuous Learning] is an open problem crucial to developing truly intelligent agents.").

[162] Everton L. Alexio et al., *Catastrophic Forgetting in Deep Learning: a Comprehensive Taxonomy*, ARXIV (Dec. 16, 2023), https://arxiv.org/abs/2312.10549.

[163] *GPT-4 System Card*, OPENAI, https://cdn.openai.com/papers/gpt-4-system-card.pdf (last accessed Jan. 20, 2026).

the state goals of China.[164] Thus, DeepSeek has no knowledge of the atrocities of Tiananmen Square or the nuanced politics behind an independent and free Taiwan.[165] Either way, the models are aligned with the beliefs and values of their creators.

Human alignment is often brittle and even breaks when switching to "low-resource" languages.[166] There are other prompt injection hacks and jailbreaking methods to get the models to ignore their alignment. For example, do anything now (DAN) is a prompting technique that allows users to remove the safeguards on GPT.[167] As the name suggests, the model would provide information on whatever the user wanted including illicit topics. When the safeguards break, the models can be used in nefarious ways such as generating racist and other harmful content.

Human alignment isn't a prerequisite for artificial intelligence which makes the question of its failures ethically grey. On Hugging Face, a repository of model weights and data, many companies provide the weights after pre-training only.[168] Thus, these models are not aligned and allow for users to finetune them for any task they desire. Human alignment is an add-on that many in tech want to placate the concerns of the broader public.[169] However, the models themselves are not particularly advertised as human aligned, at least to the general consumer (besides, perhaps, Anthropic's models).

viii. Classification Errors

We have thus far left out discussion of the most obvious types of errors from machine learning models. At their simplest, machine learning is an input and output problem. The model takes inputs from the real world and outputs a categorization, policy, or answer (among others). With the abundance of data,

---

[164] *See The Alignment Question in the Age of DeepSeek*, ALINIA, (Jan. 31, 2025), https://alinia.ai/the-alignment-question-in-the-age-of-deepseek/.

[165]Donna Lu, *We Tried Out DeepSeek. It Worked Well, Until We Asked it About Tiananmen Square and Tiawan,* GUARDIAN (January 28, 2025, 2:07 EST), https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan

[166] By low-resource, we mean languages where the amount of training data is quite small compared to English, Spanish, French, etc. *See* Yue Deng et al., *Multilingual Jailbreak Challenges in Large Language Models*, ARXIV (Oct. 10, 2023), https://arxiv.org/abs/2310.06474

[167] Sander Schulhoff, *DAN (Do anything now),* LEARN PROMPTING, (Mar. 25, 2025), https://learnprompting.org/docs/prompt_hacking/offensive_measures/dan.

[168] *Using Pretrained Models*, HUGGING FACE, https://huggingface.co/learn/llm-course/en/chapter4/2 (last accessed January 29, 2026).

[169] Even Grok at xAI has some level of human alignment using reinforcement learning, despite the concerns from Elon Musk about "woke" models. See Will Knight, *Elon Musk Signals 'Woke AI' as a Trump Administration Target*, WIRED AI Lab, (October 30, 2024), https://link.wired.com/public/37276362.

models have, on many problems, approached or surpassed "human accuracy"[170] levels.[171] As often touted in the media, machine learning models frequently outperform expert clinicians in diagnosing cancer when given imaging of a patient. For example, a meta-analysis suggests that computer models are comparable to dermatologists on identifying melanoma and outperform non-domain clinicians.[172] However, even these models that can surpass human experts will still make mistakes, just as humans would. There are obscure corner cases, latent causes, and input noise that make any classification problem inherently difficult. It is unrealistic to assume that an AI model will ever be able to adequately handle every input given the stochasticity of the real world. An AI model that outperforms a domain expert but still errs is a valuable addition to society. Thus, these types of errors are a good fit with traditional liability standards that apply when a doctor misdiagnoses a patient based on the available evidence.[173]

Hallucinations represent another variant of traditional error in large language models where the LLM will "invent" information in their responses.[174] However, if we consider LLMs (at some level) as input and output machines, then a hallucination represents an error in output tokens. We believe that grouping hallucinations with classification errors makes the most sense from a

---

[170] Surpassing human accuracy is a tricky metric to measure since the datasets are created by humans. Often, a human accuracy number refers to the expected amount of agreement between two human experts.

[171] One interesting consideration is how to handle artificial intelligence models that surpass human experts in accuracy. *See* sources cited *supra* note 86-90 and accompanying text. Many in Silicon Valley are most excited for the potential of artificial agents that can replace the human worker. Companies like Waymo have effectively created a product to remove humans from an entire industry (in their case, the taxi industry). The promise of self-driving cars is that they will exceed the safety of human operators. Let's now consider a city with two taxi companies: Waymo and TaxiCorp. Waymo has entirely self-driving cars whereas TaxiCorp has human operators who are independent contractors. Both companies have one thousand units operating at a given time. Evaluated over several years and hundreds of thousands of miles driven the driverless car has a tenth of the serious car accidents as the human operated one. When the TaxiCorp car gets into a serious accident, the liability falls on the human driver. When the Waymo car gets into a serious accident, the liability falls entirely on the corporation that created the code and vehicle involved, since there is no driver to assign blame. *See supra* Section I.B. If there are fifty serious TaxiCorp accidents and five serious Waymo accidents, Waymo will have a higher liability since they have no drivers despite having a significantly safer product. It's possible that the differential will be sufficiently high such that Waymo will be able to insure against this problem and still be more profitable than TaxiCorp, but there is still a disparity.

[172] Maria Paz Salinas et al., *A Systematic Review and Meta-Analysis of Artificial Intelligence Versus Clinicians for Skin Cancer Diagnosis*, NPJ DIGIT. MED. 1, 21 (May 14, 2024) https://www.nature.com/articles/s41746-024-01103-x#Sec29.

[173] *See* ABBOTT, *supra* note 2, at 61 (identifying this continuity though noting that integrated analysis and data collection systems blend elements of products and professionals).

[174] Adam Tauman Kalai et al., *Why Language Models Hallucinate*, ARXIV (Sept. 4, 2025), https://arxiv.org/pdf/2509.04664.

legal perspective as well. Fundamentally, we queried a model, and the model returned an incorrect prediction. The incorrectness is not from a change in data distribution or environmental context but rather an artefact of the model weights which favored an incorrect statement. Thus, we group hallucinations with other classification errors.

B. AI Failures in the Real World

i. Environmental Context and AVs

The environmental context can greatly change the liability landscape during AI failure scenarios. To expand on this, we consider the problems that arise from AVs and note that, as of the date of this writing, there are no truly self-driving cars on the market. The Society of Automotive Engineers has produced six levels of varying automation describing the capabilities of semi-autonomous vehicles. A level zero car requires full human intervention whereas a level five car requires none.[175] "Fully" autonomous taxi services such as Waymo, LLC achieve level four ratings.[176] These vehicles do not require human operators but only work under severely constrained settings. For example, a Waymo taxi can navigate select mapped cities but will fail in open environments. Oftentimes, the mapping of these cities includes high precision 3D "LIDAR" scans.[177] Even still, despite the level four rating, several lawsuits against the company indicate failures in these acceptable environments. In a recent incident, a Waymo taxi killed a neighborhood cat.[178] Tesla's "Full Self-Driving" (FSD) mode rates as a level two autonomous vehicle, requiring constant human supervision. Their website makes note of this restriction with the caveat "(Supervised)" after every mention of "Full Self-Driving".[179] However, despite these claims, a significant amount of evidence suggests misuse of these systems by the end users. Videos abound of

---

[175] *See Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, SAE INTL. (Apr. 29, 2021), https://www.sae.org/standards/j3016_202104-taxonomy-definitions-terms-related-driving-automation-systems-road-motor-vehicles.

[176] Nick Webb et al., *Waymo's Safety Methodologies and Safety Readiness Determinations*, ARXIV (Oct. 30, 2020), https://arxiv.org/pdf/2011.00054

[177] *The Waymo Driver Handbook: How Our Highly-Detailed Maps Help Unlock New Locations for Autonomous Driving*, WAYMO (Sept. 21, 2020), https://waymo.com/blog/2020/09/the-waymo-driver-handbook-mapping

[178]Sydney Johnson, *San Francisco Supervisor Calls for Robotaxi Reform After Waymo Kills Neighborhood Cat*, KQED (Nov. 4, 2025), https://www.kqed.org/news/12062777/san-francisco-supervisor-calls-for-robotaxi-reform-after-waymo-kills-neighborhood-cat

[179] *Full Self-Driving*, TESLA, https://www.tesla.com/fsd (last accessed Jan. 10, 2026)

passengers in the back seat of a Tesla with no driver to monitor the AI system.[180] Unsurprisingly, accidents and near-miss collisions have followed.

Consider the following both real and hypothetical scenarios and the evolving liability landscape. A man operating a Tesla in FSD mode in ideal weather conditions hits a seventeen-year-old student getting off a school bus. The Tesla appears to not recognize the flashing lights and stop sign on the school bus and attempts to pass the bus as it lets off passengers.[181] A woman turns on FSD mode on her commute back from work at dusk during a snowstorm. The white-out conditions and Tesla's camera-only approach to self-driving cause the car to ignore a stopped tractor trailer at a stoplight. Others have noted failures of FSD mode in winter conditions.[182] A YouTube content creator turns on FSD mode on a windy country road and immediately removes himself to the backseat.[183] The markings on the road eventually fade and the car drifts into the oncoming lane, hitting a car around a bend. In the first scenario, the Tesla clearly violates a traffic law. Although the man holds some liability for not intervening in the moment, a user should reasonably expect that the FSD car should notice the stopped school bus and abide by the rules of the road. In the second scenario, the woman should quite frankly not be using self-driving mode in such conditions.[184] In the third scenario, the driver is majority at fault for a clear abuse of the self-driving capabilities of the Tesla vehicle.[185]

ii. General versus Specialized Algorithms

Goodfellow *et* al. introduced Generative Adversarial Networks (GANs) in a seminal paper at the Neural Information Processing Systems conference in

---

[180]Tim Fitzsimons, *Tesla driver slept as car was going over 80 mph on Autopilot, Wisconsin officials say*, NBC NEWS (May 18, 2021, 5:35 PM EDT), https://www.nbcnews.com/news/us-news/tesla-driver-slept-car-was-going-over-80-mph-autopilot-n1267805

[181] *Feds Investigating Tesla that Hit Student Leaving a School Bus*, CBS NEWS (Apr. 7, 2023, 5:42 PM EDT), https://www.cbsnews.com/news/tesla-car-crash-nhtsa-school-bus/

[182] Mack Hogan, *"Testing" a "Full Self-Driving" Tesla in Detroit Snow Looks Reckless and Embarrassing*, ROAD TRACK (Jan. 3, 2023, 4:11 PM EST), https://www.roadandtrack.com/news/a42384858/testing-a-full-self-driving-tesla-in-detroit-snow-looks-reckless-and-embarrassing/

[183] Several individuals have been arrested previously for sitting in the backseat while using Tesla's self-driving mode. *See* Johnny Diaz, *Man Riding in Driverless Tesla is Arrested in California*, N.Y. TIMES (May 12, 2021), https://www.nytimes.com/2021/05/12/us/california-tesla-backseat-driver.html

[184] See the *Autopilot: Limitations and Warnings*, MODEL 3 OWNER'S MANUAL, https://www.tesla.com/ownersmanual/model3/en_us/ (last accessed: January 26, 2026) ("Visibility is critical for Full Self-Driving (Supervised) to operate. Low visibility, such as low light or poor weather conditions (rain, snow, direct sun, fog, etc.) can significantly degrade performance.").

[185] Perhaps with the caveat that Teslas do have a weight sensor on the front seat and should thus stop forward progress of a vehicle in such instances.

2014.[186] Two neural networks, a generator and a discriminator, "compete" against each other in a game. For the purposes of this discussion, we will focus on the generation of images and videos but note that the technology extends to other modalities such as text. The generator's goal is to produce photorealistic images. As the name implies, the discriminator's goal is to determine which images are artificially versus naturally created. In this iterative game, both the generative and discriminative models improve by producing more natural-looking images and deciphering between real and artificial images, respectively. Alternatively, but with a similar end goal, diffusion models produce photorealistic images by sampling from random noise.[187] Companies such as Stability AI provide convenient web interfaces for individuals to play with these types of generative AI and produce images from various prompts.[188] Its primary offering, Stable Diffusion, is a multi-modal model that converts natural language into stunning images. Other major players have similarly followed suit with one able to query, for example, OpenAI's tools on (and to create) images. These image generation tools allow one to provide the type of image (photorealistic, renaissance, baroque, cartoon, etc.) as well as the subject, among other parameters.

Although these image generators can greatly facilitate the creation of figures, digital advertisements, and renderings for creative types, they can be used for more nefarious purposes. Deepfakes leverage these generative image technologies to create realistic but fake renderings of celebrities, politicians, and even everyday citizens.[189] These deepfakes can be used as "evidence" in misinformation campaigns.[190] In some of the most disgusting abuses of this technology, (ab)users have generated portrayals of famous actors and actresses in pornographic scenes.[191] Some states have added deepfakes to existing revenge pornography laws to provide additional protections for their citizens.[192]

---

[186] Ian J. Goodfellow et al., *Generative Adversarial Networks*, PROCEEDINGS OF THE 28TH NEURAL INFORMATION PROCESSING SYSTEMS 2672 (2014), https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html

[187] Sohl-Dickstein *et al.*, *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*, 37 PROCEEDINGS MACHINE LEARNING RESEARCH 2256 (2015)

[188] *See Stable Diffusion 3.5*, STABILITY.AI, https://stability.ai/stable-image (last accessed Jan. 30, 2026).

[189] Gan Pei et al., *Deepfake Generation and Detection: A Benchmark and Survey*, ARXIV (May 26, 2024), https://arxiv.org/abs/2403.17881.

[190] *Increasing Threat of DeepFake Identities*, DEP'T HOMELAND SEC. 19-20 (Sept. 29, 2023), https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

[191] *Id.* at 5.

[192] Karen Hao, *Deepfake Porn is Ruining Women's Lives. Now The Law May Finally Ban It,* MIT TECH. REV. (February 12, 2021), https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/.

It seems obvious that tools like Stable Diffusion that specialize in image generation should have internal safeguards from such abuses of their system. Indeed, if you ask OpenAI to generate an image for a paper on tort liability and AI, the model will produce images of generic people (and robots). When one specifically requests an image that includes the likeness of the current Supreme Court, the model abstains citing that it cannot produce likenesses of real, identifiable individuals. Note, however, the model does produce the likeness of historical figures (in our case, first chief justice of the United States, John Jay). It is unclear the threshold for historical figures—Founding Fathers seem obviously historical whereas a more recently deceased politician, such as George H.W. Bush, seems less so. That aside, these mainstream image generators ought to, and do, have safeguards to prevent the malicious abuse of their system with deepfakes. Liability for the creation of deepfakes using these tools should generally fall on the companies that produce the tool. The use of generative image technology for deepfakes is common knowledge and the providers of these tools must safeguard against the abuse of the tools. We refer to these as specialized tools for the image generation task.

However, recall that these generative tools are building on existing algorithms in the academic literature. Furthermore, OpenAI and others have trained their models on a large corpus of academic literature and source code on GitHub and other code repositories. One could simply query GPT-5 to produce code that implements a diffusion model. Skeptics will (rightfully) point out that having a model architecture is significantly different than having a working model. The training data (and hyperparameters) are the secret sauce that differentiate a small academic/research model from a full-scale customer-ready tool. However, when it comes to deepfakes, one doesn't necessarily need the best generated image to provide sufficient harm to the community (in the case of misinformation) or to an individual (in the case of revenge pornography). Additionally, many research papers provide their model weights for free to the broader community.[193] In these scenarios, who is liable for the creation and proliferation of deepfakes? Surely not the general-purpose AI tool that provided the algorithmic framework and model structure for the wrongdoer or the research scientist who uploaded weights based on existing requirements from the National Science Foundation. Instead, when using a general-purpose tool to construct an algorithm for malicious purposes, the liability would shift to the (ab)user.

We note that this specific scenario relies on the creation of a tool that has both benign and malicious use cases. Other uses of general knowledge tools are clearer cut. An individual probing Claude to find (zero-day) vulnerabilities in a

[193]Despite the potential for misuse, this is considered a net positive for science in a world with increasing skepticism of reproducibility of results! In fact, some funding agencies now require researchers to provide their code and weights for the public on the condition of receiving funds.

binary or source code is likely trying to abuse the model's capabilities. These aren't hypotheticals—state actors have recently been caught using large language models to streamline their malicious activity against the United States government and major businesses.[194] The liability landscape must also match the gradients of these abuses.

iii. Digital Assistants

The most bullish Silicon Valley venture capitalists envision a future where autonomous agents replace individuals in a corporate setting. However, these corporate autonomous workers require a unique liability discussion that is highly dependent on the producer's marketing and product offering. Already, some corporations and governing bodies are replacing human operators with autonomous agents with mixed results. Tessa was a chatbot created to aid the National Eating Disorders Association (NEDA) field the large number of callers and users reaching out for emergency help (70,000 in 2022 alone). Unfortunately, Tessa quickly provided those in need with disastrous advice, focusing on weight loss techniques with an emphasis on results in pounds and not in overall health. This is the absolute worst-case scenario for many in need who for years struggled with body confidence.[195] Tessa was created by health researchers specifically for this use case and sold to replace the team of human operators previously working the helpline.[196] Thus, its failure falls on the developers of the technology.

Consider an alternate world where researchers (with funding from NEDA) did not invest time and money into building Tessa. Instead, buoyed by the successes of Chat-GPT on academic benchmarks and excited by the possibilities for LLM personas through prompt engineering, NEDA created a web interface using OpenAI's Application Programming Interface (API) to field questions from callers to the helpline.[197] This chatbot similarly fails spectacularly, providing weight loss regiments to those looking for body acceptance. In this

---

[194] *Same Targets, New Playbooks: East Asia Threat Actors Employ Unique Methods*, MICROSOFT THREAT INTELLIGENCE (April 4, 2024), https://www.microsoft.com/en-us/security/security-insider/threat-landscape/east-asia-threat-actors-employ-unique-methods.

[195] Kate Wells, *An Eating Disorders Chatbot Offered Dieting Advice, Raising Fears About AI in Health*, NPR (June 9, 2023, 6:59 AM ET), https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea

[196] Kate Wells, *National Eating Disorders Association Phases Out Human Helpline, Pivots to Chatbot*, NPR (May 31, 2023, 5:08 PM ET), https://www.npr.org/sections/health-shots/2023/05/31/1179244569/national-eating-disorders-association-phases-out-human-helpline-pivots-to-chatbo

[197] An application programming interface (API) allows a user to interact with an underlying system. In this instance, the OpenAI API allows a user to query an OpenAI model and provide additional context to the model to modify the outputs and persona through prompt engineering.

instance, OpenAI made no promises to the end customer (NEDA) on the usability of its API for such a purpose. The liability would fall solely on NEDA for its misuse of the OpenAI tool. There is also no real way for OpenAI to know how NEDA will use its tool. Even Tessa provided numbers-based approaches to the questions asked! The problem is that a numbers-based approach is the exact wrong mindset for a chatbot of an eating disorder helpline. Although we use a hypothetical in this second scenario, it mirrors the expected trajectory of adoption of AI agents into corporate workflows. Sam Altman has stated that he believes an AI agent will "soon" replace him as CEO of OpenAI.[198] (We do note, however, that Sam Altman has a strong financial incentive to overhype the future capabilities of agents in replacing humans in corporate settings.)

Currently, some companies are focusing on domain-specific AI agents for the professional class. One such example is Harvey, a digital assistant for lawyers.[199] These domain-specific agents provide a unique legal landscape since they are advertised for the professional class. As an end-user, a lawyer can significantly improve productivity using Harvey when analyzing existing proprietary documents (through a RAG-based model)[200] or finding relevant court cases (open-ended search). Despite the intended use case, it is likely that some companies will use Harvey (or equivalents) as *their* legal counsel (and in fact some overly online tech founders already admit to using OpenAI products for this, perhaps prompting the OpenAI to clarify that it would not offer personalized legal or medical advice, until its decision to launch a health product).[201] In these instances, the highly-tailored AI assistants should abstain from providing legal advice as if they were legal counsel. Harvey, as the assistant, ought to know what falls within acceptable use. A trained lawyer will provide significantly different queries than a general user looking to replace their legal fees. Furthermore, an AI that is designed for the professional class should, with limited overhead, be able to confirm that the user actually belongs to the professional class (e.g., is a lawyer in good standing with the state Bar).

---

[198] *Sam Altman says AI May Replace Him as OpenAI CEO, He Will do Farming When It Happens,* INDIA TODAY (Nov. 6, 2025, 13:59 IST), https://www.indiatoday.in/technology/news/story/sam-altman-says-ai-could-soon-replace-him-as-openai-ceo-he-would-like-to-be-a-farmer-when-it-happens-2814498-2025-11-06

[199] *See Company*, HARVEY, https://www.harvey.ai/company (last accessed Jan. 30, 2026).

[200] Retrieval-augmented generation (RAG) is a technique where context from a document (or knowledge) base is provided to the LLM to help ground the answer and reduce hallucinations.

[201] Elianna Lev, *ChatGPT Users Can't Use Service for Tailored Legal and Medical Advice, OpenAI Says,* CTV NEWS (Nov. 5, 2025, 4:12 PM EST), https://www.ctvnews.ca/sci-tech/article/chatgpt-users-cant-use-service-for-tailored-legal-and-medical-advice-openai-says/. Though later OpenAI launched such a product for health. *See* source cited *supra* note 44.

C. Academic Progress and Closing the Foreseeability Gap

Section III.A discusses unique types of errors that occur when introducing an artificial intelligence agent into an open world where the sterile training environment might deviate from reality. A large amount of research explores how to address the challenges when models make the leap from lab to shelf. Expectedly, the "safeguarding" research trails the cutting-edge products on the market.[202] It is likely that some of the problems enumerated here will have accepted mitigations (perhaps even in the near future). As these new methods become more readily accessible, it would be expected that commercially available AI tools would utilize these tools in the same way that automobiles built today have seatbelts and airbags. As such, the liability would shift towards the producer for ignoring the advancements of such algorithms as AI would become more like a well-understood product and the duty of care of its developers would become clearer.

IV. OVERCOMING (UN)FORESEEABILITY

A. Foreseeability: Why, not How

With the above in mind this section turns to our proposal. The main thrust of our idea is that the proper way to resolve the foreseeability gap is to neither adopt a very high level of generality nor a low level one. Instead, the proper response is to take a middle ground approach and focus on *why* the unpredictable error occurred using the categories outlined above.

For example, when confronted with an "unpredictable" error we can ask whether, for example, it was the result of catastrophic forgetting or another known reason for failure. If so, then the error *may* result in developer liability. In particular, once we can trace the error we can ask questions about whether appropriate mitigation steps were taken and whether liability should shift to some other actor in the chain.[203]

Our approach thus sits somewhere between negligence and strict liability. Unlike traditional negligence, our notion of foreseeability is somewhat broader and relaxes questions of proximate cause. For example, consider a driver who is

---

[202] Hallucinations were not in the academic vernacular until the proliferation of chat bots that mimicked natural language. Now, identifying and mitigating hallucinations is an active area of research with multiple NeurIPS papers exploring the area.

[203] Though we discuss notice next in *infra* Section IV.B, we take no firm position here on the question of what mitigation steps would be sufficient to provide a defense to an accident caused by a traceable error. Failure to take well-known and easy to implement steps strikes us as an easy case for traditional legal rules, but the contours of more difficult cases is a more difficult problem. *See supra* Section I.D (discussing negligence and, in particular, professional standards of negligence).

speeding and causes an unusual accident. Our proposal would hold the driver liable on the grounds that speeding is a known *way* in which accidents occur even if the precise nature of the accident was unforeseeable. Similarly, as discussed, strict liability for wild animals holds keeper strictly liable provided the accident is of a *type* typical of the animal.[204] While the mechanism by which the tiger might maim an individual may be unforeseen, the type of harm matters for liability. Our proposal is focused on the mechanism of harm and thus departs from strict liability.

This proposal we think balances several important considerations. First, it compensates victims of AI models by limiting the foreseeability gap. While unforeseeable accidents due to emergent behavior will sometimes go uncompensated, our proposal gives plaintiffs tools to address unforeseeable accidents. Second, by focusing on known mechanisms of failure, we think our proposal is also fair to developers of these models. Developers can take steps to limit failures due to these reasons even if it will be difficult to eliminate all errors. Because we believe Artificial Intelligence has the potential to benefit others, we think avoiding an overly punitive approach is helpful.

Three further clarifications: First, as noted, our proposal is not intended to displace other doctrines of liability. In particular, we think the various proposals discussed above indicate that it is important for the treatment of AI to be "continuous at the boundary." By this, we mean that where an AI system acts like an ordinary object of legal regulation, we should treat it using traditional rules. An AI-enabled product that fails in a way that can be handled by traditional product liability rules should be treated using those rules.[205] Similarly, an artificial employee that harms a third party due to negligent supervision should subject the company to traditional vicarious liability.[206] Continuity in these cases both promotes parity between AI and non-AI systems and also reflects our respect for the wisdom of these traditional rules. As discussed below, it also has useful implications for how our rule will fit within the legal landscape as we go forward.

Second, in some cases these disputes are best handled by contract, especially between sophisticated actors. Without wading into the waters of debates about consumer contracts, in some cases industry actors are best placed to allocate risk between themselves. Because one of the paradigm cases that motivates us is harm to third parties caused by AI systems deployed by a user not involved in developing the underlying technology, we provide a tort solution here. Still, contract law has a role to play in departing from the defaults to best fit the business context.

---

[204] *See supra* notes 106-112 and accompanying text.
[205] *See supra* Section I.A; *see also* supra notes 101-102 and accompanying text.
[206] *See supra* notes 103-105 and accompanying text.

Third, as we discuss in the next Section, our proposal is not that companies are 100% liable if an accident results from a known mechanism. Rather, we see this as a threshold condition for considering liability for the developer. Defenses or other considerations may counsel in favor of limiting liability in some cases. And, critically, in some cases responsibility should shift partially or entirely to the end user based on notice. We turn to this topic next.

B. Notice and the Distribution of Responsibility

Turning to how to evaluate users of AI whether they themselves are injured or deploy systems that injure third parties, we think here that the proper allocation of liability turns on questions of notice. We have in mind three types of failures: First, specific known failures of the model (e.g., the model will fail if prompted in this way); second, situations in which the model is likely to fail (e.g., the self-driving system will fail in heavy precipitation); and, third, the general knowledge that models can fail due to emergent behavior.

For the first type of failure, we think traditional products liability standards suffice. Where a model is known to fail in a particular context in a particular way we can ask whether there's a failure to warn, some sort of manufacturing defect, or a design defect in the usual way. Basically, if the error can be easily eliminated (or eliminated in a cost-justified way) then the manufacturer should do so.[207] If not, under appropriate circumstances a warning should be given.

For more specialized actors, moreover, a proper answer might be to ask whether they are on inquiry notice about the particular error. For AI systems this would essentially function akin to a negligence supervision/hiring claim. Thus, a doctor or hospital deploying a medical AI likely should have a duty to ask reasonable questions about how and why the model might fail (leaving aside that the doctor has a duty to provide medical advice akin to a reasonable doctor). Similarly, a corporation "hiring" a digital employee should have a duty to take basic care to understand the limitations of the model. The superior expertise of the specialist as well as (in some cases) the special trust given to these actors, we think, generates a higher duty to protect third parties.

The same general logic applies where to cases in which the artificial intelligence system is known to be unreliable in a particular context. Take, for example, self-driving vehicles in highly inclement weather.[208] The product is known to fail or be unreliable in such situations and thus we can use traditional product liability to distribute responsibility. We can ask whether proper warnings were given that the product could be dangerous if used in this way even if the product cannot be redesigned to fix the problem. And again, for specialized actors we can ask whether they were on inquiry notice about the use.

---

[207] *See supra* Section I.A.

[208] *See supra* Section III.B.i.

Of note, we contemplate that some errors may be expected yet unexplainable. Thus, a problem may be known, and users may be on actual notice or inquiry notice about the problem, yet researchers may be unable to trace back the source of the error. In this case, we think application of traditional negligence principles leads to the conclusion that in the appropriate case the user has a duty to protect others from the error since it is sufficiently foreseeable. Similarly, applying traditional failure to warn principles, an AI developer may have a duty to provide a warning about such an error even if it is not traceable to a particular mechanism.[209] This generates the conclusion in the bottom right of Figure 2. In either case, we see these cases as governed by traditional principles and thus outside of our gap-filling intervention.

These principles fit with the previous Section because known ways in which the model fails can help shape what warnings should be provided. Moreover, leading artificial intelligence companies are known to engage in extensive testing of their models in order to determine how they can be fooled.[210] Thus we think that information-forcing rules that encourage the companies to transfer knowledge they have about the model's failures do seem beneficial when coupled with a view that roots liability in the known ways in which models generate unpredictable results.

To be clear, we take no position on the precise allocation of responsibility between developers and users. In some cases where the model fails due to a known mechanism, proper warnings may defeat liability for the developer. In other cases, it may not. Finally, comparative fault principles might be appropriately applied to distribute responsibility among actors. Finally, we imagine that just as in traditional tort law there will be cases where the error is so odd that tort law ought to let the losses lie where they will. These two considerations produce the figure in the introduction.

---

[209] *See supra* notes 35-38 and accompanying text.

[210] *See* Kim Martineau, *What is Red Teaming for Generative AI*, IBM (Apr. 11, 2024), https://research.ibm.com/blog/what-is-red-teaming-gen-AI ("Red teaming for generative AI involves provoking the model to say or do things it was explicitly trained not to, or to surface biases unknown to its creators. When problems are exposed through red teaming, new instruction data is created to re-align the model, and strengthen its safety and security guardrails.").

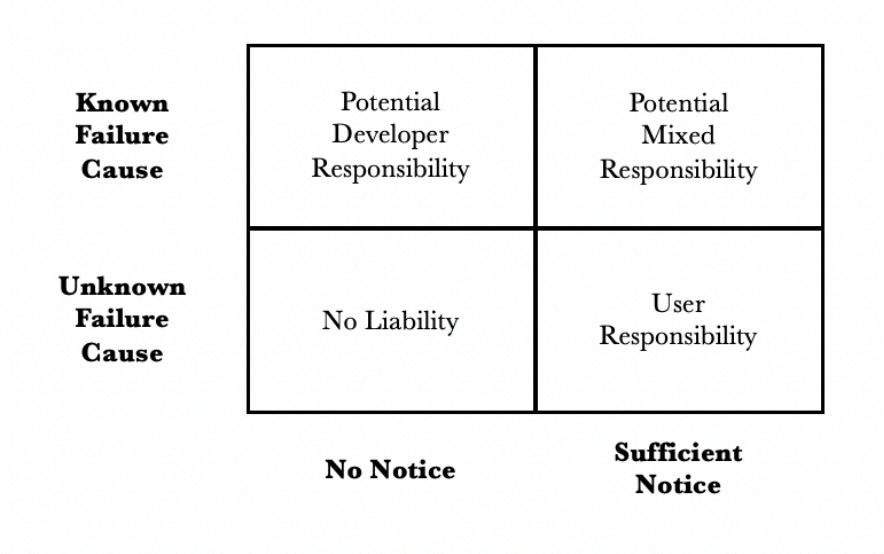

Figure 2: Liability Under the Model

In particular, where the model displays some emergent behavior that is unknown and truly unforeseeable the foreseeability gap will remain.[211] As technology and society advance however, we believe our framework helps answer some of these questions. The next Section tackles this problem.

C. Going Forward and the Need for Future-Proofing

What happens as technology and society change and advance? We believe our model has several advantages when it comes to dealing with change over time. In particular, it is technology agnostic, designed to fit with changes to the applicability of traditional doctrines, and is able to tackle changes to the frontiers of knowledge about AI.

First, our model is not keyed to any particular AI technology such as LLMs. Provided that an AI is designed to take in inputs and predict a "best" output, we can ask questions about why the model may fail to do so. Thus, while social experience with a particular type of model (such as LLMs or AVs) may narrow the foreseeability gap as to that technology, new AI technologies may function differently.[212] As so the gap may once again widen as technology advances. We

---

[211] Though potentially in some cases we can imagine the user of the AI being liable if they deploy the system in a careless way that creates a risk of harm.

[212] For example, spiking neural networks (SNNs) are an emerging machine learning algorithmic paradigm that attempts to mirror the spiking mechanism in biological neural networks. One of the benefits of this approach is that the models can (but are not required to) run on neuromorphic hardware, brain-inspired chips that are highly energy efficient, especially compared to the currently used graphical processing units (GPUs). The fact that these SNNs rely on spiking thresholds though means that some inputs are not strong enough to trigger a response. Thus,

thus believe it is an advantage of our proposal that it can be adapted over time as new technologies come onto the market.

Second and relatedly, we note that our proposal is flexible to dynamics that will alter the nature of the foreseeability gap. One, which we noted earlier, is simply that as people become more and more experienced with artificial systems the nature of foreseeability will change. In particular, certain types of accidents will become foreseeable over time once these systems are in ordinary use and thus traditional tort principles will have an easier time applying. In such cases, it makes sense to apply those principles. Because our approach is intended to fill a gap and not displace traditional tort law, we have no problem with such changes.

Third and finally, our proposal does not assume the state-of-the-art in AI interpretability and understanding will remain fixed. Over time, more mechanisms for AI failures may come to be understood and thus will generate liability according to our proposal. This is an advantage because in practice our proposal gives companies and users incentives to fold these state-of-the-art results into their duty of care. Moreover, the expansion of the duty of the developer does not strike us as unfair. It only asks developers to take responsibility for what they can know; not to take responsibility for truly unknowable outcomes.

In sum, we think the advantage of our framework is that it is relatively flexible and future proof. As technology changes, society shifts, and research advances, all three dynamics can be accounted for by our proposal. Indeed, even if our solution is not adopted, we believe that a satisfactory solution to the problem of artificial intelligence liability ought to have these characteristics in general. Our solution's fit with these criteria is a point in its favor.

* * *

We acknowledge that this treatment is a sketch that does not work out all of the contours of how such an approach would work. We note two particular omissions at this juncture. The first is that we are silent on the procedural aspects of our approach. While we believe that problems of proof can be overcome, we recognize that courts will need to work through these problems. The second is that we note that this treatment is silent on industry specific problems that may call out for specialized solutions. Our approach here is meant as a suggestion for a high level guidepost but we realize that specific industries and business contexts may call out for specialized solutions. For example, in the case of self-driving cars, it may be the case that the sensible equilibrium is for developers to simply obtain insurance that covers the costs of accidents. Moreover, there are many

---

there may be information loss during the processing of the inputs from the environment. Unlike traditional deep neural networks where even weak signals propagate as small activation values, sub-threshold inputs in SNNs produce no spikes and transmit no information downstream. This may produce unexpected errors when deployed in the real world in the future.

different patterns of AI development, deployment, and interfaces with external parties that may call out for more specific analysis. We leave these specific considerations for future work.

CONCLUSION

AI systems have enormous potential. By "thinking" in ways that allow them to solve human problems often better than the typical person, they promise to unlock significant economic value and save lives. The issue for holding such systems liable is that the ways in which they "think" is unlike human beings and thus when they err they do so in ways that are often unforeseeable and thus outside the reach of traditional tort law. Existing literature often either admits the failure of tort law caused by what we term the "foreseeability gap" or attempts to argue that AI can be governed by an existing legal doctrine or category. But embedded in those doctrines and categories are social assumptions and understandings about how, when, and why accidents occur. While we may develop such understandings over time, we have not done so yet. What we do know is many of the reasons *why* AI fails.

In this paper, we have enumerated several ways in which artificial intelligence fails in a way that a human would not. One could argue that these errors are, in fact, foreseeable as evidenced by the fact that we discuss the errors at length with numerous real-world examples. To this, we make two additional points. First, this list is non-exhaustive but more importantly ever-growing. Researchers discover every day new ways artificial intelligence fails. Second, each of these failures (outside of classification errors which follow from type errors in traditional statistics) was previously unknown, and thus unforeseeable. Before 2017, it was generally not known that deeper models were simultaneously becoming more accurate and more over-confident on wrong predictions![213] Researchers at OpenAI discovered that a robot tasked with grasping objects would simply move the table to remove the object from view and satisfy its reward function.[214] As AI continues to develop, we may discover new errors that call out for research into their root causes and potential ways to mitigate them.

This work is intended to be a first step towards an approach that leverages this work. By rooting liability in *why* the AI failed we believe we can leverage this process of discovery and research to find a level of foreseeability in unpredictable AI errors. In doing so, we believe we can balance fairness to developers while still ensuring that victims of AI accidents are compensated.

---

[213] Guo, supra note 153.

[214] Dario Amodei, Paul Christiano & Alex Ray, *Learning From Human Preferences,* OPENAI (June 13, 2017), https://openai.com/index/learning-from-human-preferences/.