Bridging the Gap Between Tort Law and Unforeseeable AI Errors

Mitchell Chervu Johnston [1] **Brian Matejek** [2]

Abstract

Artificial intelligence does not simply make errors; it makes unusual, even unforeseeable, errors. Since traditional tort law categories embed foreseeability within them, these unforeseeable errors vex attempts to fit AI harms fully into traditional doctrinal boxes. We argue that this "foreseeability gap" generates many of the hard problems of applying tort law to AI systems. We argue that the foreseeability gap can be better closed by careful attention to the way in which AI systems work. We propose a novel approach to liability rooted in the reason for the AI failure. Such an approach requires AI developers to prioritize *provenance* of their workflows. This proposal, we argue, balances fairness to developers and users of AI with the goal of compensating victims of accidents.

1. Introduction

A small drone delivery service "hires" a software engineering AI agent for schedule optimization, called Alice, from a major AI company that sells digital assistants for numerous targeted roles. Buoyed by a dramatic increase in productivity but stymied by their own financial reality, the corporation decides to re-purpose (through a small amount of fine-tuning and prompt engineering) Alice to improve their in-flight control mechanisms. However, Alice was never intended to work on this technology—that was a separate offering, named Bob. Alice uses an unsafe open source python package with a recently discovered CVE that allows adversaries to seize control of the drone. Attackers target one of the company's drones causing their client to sustain significant property damage. The client now seeks financial restitution. The drone delivery company points blame to the digital assistant manufacturer which promised a level of security competency for all agents. The AI manufacturer claims that the post-sale fine-tuning of Alice corrupted its security competency. Which party, if any, should bear the legal responsibility for the client's injuries? Did the fine-tuning erase Alice's security competency? Should Alice know of a CVE discovered after its training date? Is it even reasonable to advertise security competency if the digital assistants do not receive weight updates? The company might point to a waiver of liability included in its terms of service, but such waivers are of uncertain enforceability, and, moreover, the client is not a party to the agreement between the AI company and the drone service.

Traditional frameworks of tort liability implicitly embed various social understandings about the types of errors that others make. They assume not only that humans err but that these errors follow a regular distribution that other actors can understand. Normal academics do not typically invent citations out of whole cloth, AI "researchers" do (Warren, 2025). By now, of course, academics 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 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 (Shah et al., 2022).

Proposals that attempt to fit artificial intelligence into traditional legal categories overlook that artificial intelligence regulation is tricky, in part, because the structure of AI errors is *unusual* (Schneier & Sanders, 2025; Kannegieter, 2026). AI systems are said by some legal commentators to be a "black box" (though we think "opaque" or "of uncertain visibility" are superior descriptions of the problem), true, but so are the minds of others. Unpredictability, incomprehensibility, and oddity are the salient features of AI harms that call out for a response (Bathaee, 2017; Choi, 2023). To be sure, in some cases AI harms occur in ways that closely resemble traditional legal problems. For example, an AI employee might harm a third party in the same way as a human employee. Here, we think the law should apply traditional tools. It is the unforeseeable errors that generate the hard cases that have sparked debate.

To be clear, legal foreseeability is not a purely probabilistic assessment. It involves normative considerations and therefore cannot be conflated with predictability. Our contention that existing doctrines rely on foreseeability is, in part, a

[1]Boston College Law School [2]Computer Science Laboratory, SRI International. Correspondence to: Mitchell Chervu Johnston <johnstm@bc.edu>, Brian Matejek <brian.matejek@sri.com>.

Second Workshop on Technical AI Governance Research (TAIGR) @ ICML 2026, Seoul, South Korea. 2026.

statement about fairness and, in particular, the judgment that it is fair to ask a particular actor to take into account a potential outcome when choosing to act. What makes certain AI systems difficult to regulate using traditional foreseeability tools is the intersection of technical unpredictability, opacity, and the absence of well developed social understandings about the fair distribution of responsibility for AI harms.

However, not all AI errors are *unexplainable*. A sizable amount of academic research focuses on dissecting the symptoms and causes of AI failures such as hallucinations (Kalai et al., 2025), catastrophic forgetting (Aleixo et al., 2023), and reward hacking (Clark & Amodei, 2016), among others. The gist of our proposal is that the law should rely on these findings to adapt the concept of legal foreseeability. However, this requires attorneys representing both plaintiffs and defendants to understand the underlying causes of a failure and *to document the path from input to harm*. Plaintiffs attempting to show that the error was not truly legally unforeseeable would need to adequately prove that the error can be classified into a known error type while defendants would aim to rebut their case.

Such a legal framework would require AI developers to employ some *provenance* measures on their models. Cases of implicit model bias would require *data provenance* where one can parse the exact data used to train a given model. Companies that are currently not employing such provenance measures would now find financial and legal incentive for doing so: if a defendant wishes to shift liability by claiming "unforeseeability" or emergent behavior, the court needs at least some insight into the inner mechanisms of their AI system. For *inference provenance*, this would require, among other things, regular snapshots of parameters, maintaining random seed values, and maintaining representations of the input data.[1] From here, experts could recreate errors in a court of law similar to how a forensic tire track expert can digitally recreate an accident for a jury.

This notion of provenance is closely related to the ideas of traceability and audit trails discussed in prior literature in the trustworthy AI ecosystem (Brundage et al., 2020; Li & Goel, 2025). Fundamentally, the courts and regulators would like to trace the decision making process of an AI system, similar to how the National Transportation Safety Board might analyze an airplane's "black box" after an accident. We understand that such a shift for AI developers, particularly the AI giants with massive existing infrastructure, cannot happen overnight. Therefore, we envision a scenario where there are varying levels of provenance for different components of an AI system. Gradually the scope of provenance increases until, ideally, there is sufficient provenance over an entire AI system for a court to trace from input to output.

Our approach is based in US law and US tort law in particular. Tort law is a branch of judge-made law that governs non-contractual claims of harm between parties. Its principles are highly general and thus it has proven to often be quite flexible in its ability to handle technological change. That said, we acknowledge that the statutory solutions being developed both within the US and outside (e.g., the EU AI Act) have much to recommend to them. In particular, because tort law focuses on what happens after an accident occurs, it is less able to provide detailed *ex ante* guidance to regulated parties. That said, to the extent the principles and ideas articulated in this position paper are attractive, they could be incorporated into a statutory scheme to supplement other interventions.

We make two main points in this position paper. First, that the current legal framework is inadequate to handle AI technologies in *some failure cases*. Where existing doctrine reasonably apply, the courts can default to that existing doctrine (e.g., when a digital assistant produces the same error as their human counterpart). However, when AI fails in unforeseeable ways, the liability landscapes changes. Second, AI developers need to adjust their practices based on this legal paradigm. If unforeseeability is an important legal defense for an AI manufacturer, lawyers need to have a mechanism in place to prove (or disprove) the truly novel nature of the error in court. Furthermore, as researchers study these systems and identify new failure cases, and the courts and media publicize their downstream harms, more things will fall under the umbrella of foreseeable errors.

2. Previous Proposals

As AI systems become more agentic, they blur the lines between software and human action. Nevertheless, in some cases, AI errors (or the errors of their developers or deployers) and the harms they cause will neatly fit into one box of the other and thus traditional legal tools can be applied. If an employer negligently supervises an "AI employee" and a harm results that is closely analogous to what would have occurred if a human was doing the same job, then it may make sense to apply traditional legal solutions to the problem.

Scholars have proposed various ways to apply and expand traditional categories to the hard problems of AI. This Section (very briefly) discusses those proposals and what we see as their limitations in addressing unforeseeable errors.

2.1. Strict Products Liability

Strict products liability is a bundle of three related doctrines courts developed to address harms caused by commercial

[1]There are other significant challenges with recreating outputs from inputs, mostly stemming from the non-associativity of floating point arithmetic (Ingonyama, 2024). PyTorch provides a short list of best-practices from trying to reduce the non-determinism during inference (PyTorch Contributors).

products. First, *failure to warn* claims require companies to provide warnings to consumers where those warnings can help prevent injury. Second, *manufacturing defect* claims hold manufacters liable where the products departs from its intended design in a way that causes injury. Third and finally, *design defect* claims generate liability for manufacturers if plaintiff can prove that the design of the product itself is somehow defective and an alternate design could have prevented injury (American Law Institute, 1998, § 2).

Of these three types of claims the most important and most difficult question is how to apply design defect claims to AI systems. The modern trend in design defect cases is to require a reasonable alternative design for the product. Moreover, plaintiff must show that the alternative design's reduction of risk is not outweighed by a reduction in the utility of the product. While obvious cases may present themselves in which system designers ignored clearly safer designs, outside of these easy situations it is unclear how to apply this criteria. (Indeed, if plaintiff is able to put forward an AI system which reduces risk without compromising utility, he or she may be well advised to commercialize the insight) (Vladeck, 2014).

2.2. Strict Liability

Another proposal would be apply true strict liability to developers and hold them responsible for all harms caused by their models with few limits.[2] Strict liability of this sort has two flaws: one conceptual, the other practical. On the conceptual side, strict liability of this sort abandons the requirement of fault on behalf of the defendant. Accordingly, it makes sense that traditionally it has been restricted to activities that are so dangerous or unusual that we might question the pro-sociality of engaging in them. AI development and deployment seems a poor fit for this criteria. Second, even if strict liability for AI harms had a stronger conceptual footing, it faces a critical practical problem: If strict liability leads developers or deployers to expect massive losses if a catastrophic error occurs it may lead developers to *underinvest* in safety (Weil, 2024). For example, consider a developer who judges that the probability of a catastrophic risk is 2% but with adequate investment in safety it can be reduce to 1%. Under a no-fault scheme of liability the developer will be responsible if the risk materializes independent of investment in safety. If realization of the risk will bankrupt the company in either case, there is a reduced incentive to take precautions *even if* the precautions are cost-justified when considered in relation to the harm to third-parties.

[2] True strict liability is applied for harms caused by the keeping of wild animals and so-called abnormally dangerous activities and requires only that the harm be of the type that makes the activity dangerous (e.g., the tiger must harm the plaintiff in a tiger-like way) and that legal standards of causation be met.

2.3. 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 agentic systems with greater capabilities. The question then becomes how to hold companies liable for torts or other accidents committed by these employed algorithms (Diamantis, 2022; Lior, 2019). Traditionally, employers are held liable for the torts of their employees (independent contractors have their own rules) if the employee acts within the scope of their employment for the benefit of the employer. To sweep in artificial employees, some adjustments to this doctrine are likely required, though attractive proposals exist (Diamantis, 2022).

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. Where artificial intelligence systems err in different ways than human employees, 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.

2.4. 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. If an actor exercises ordinary care, they will not be liable. Moreover, even if they fail to act with ordinary care, negligence liability limits liability based on the foreseeability of accidents. Again, we agree at a high level 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 (Schneier & Sanders, 2025; Vladeck, 2014). One way to address questions of ordinary care is to adopt a professional care standard (Choi, 2023; Sharma, 2024). But given the novelty of the field and the fast-moving development of AI systems, the foundation for this approach is uncertain (Choi, 2023).

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. But this raises the question of what types of accidents are foreseeable and who can foresee them (Lior, 2019). The errors we are concerned with are *unpredictable* and *unusual* and thus raise a question of how to adapt this traditional doctrinal tool.

2.5. No Fault/No Liability Approaches

The final category of proposals would absolve AI developers of fault and instead introduce a no-fault compensation scheme, akin to a form of mandatory insurance (Vladeck, 2014). 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. First, one could worry that regulation (or regulation by lawsuit) will chill innovation (Ball & Rozenshtein, 2024). If so, sweeping away this risk and instead asking companies (or the state) to pay for the harms they cause could be justified given the potential for economic uplift.

Second, one could simply take position that the hard problems of AI liability are indeed the result of risks that are unforeseeable (Vladeck, 2014). If so, it is unfair to hold anyone liable for them under a fault-based regime. The better response is to simply conclude that someone must act as an insurer for those harmed by AI and work out who is best suited to pay (Yoshikawa, 2018; Lior, 2019). The scope of insurance coverage will vary based on the scope of the problem and who should pay may depend on the situation, but if attribution of blame cannot be done in a morally attractive way, insurance seems to be the best approach.

3. Tort Law and the Problems of AI

3.1. The Human-Fallibility Assumptions of Tort Law

In Section 2, we discussed how various proposals for AI liability struggle to adapt to the nature of AI Errors. The issue, we believe, is that the law allocates responsibility based on assumptions about how and when humans err and social knowledge of these assumptions. In some cases, the law asks us to be aware of our own foibles and we are held liable based on when we fall short of reasonable standards. Things become more controversial when we are held liable for the errors of others. Anchoring the law in social understandings of how others err helps ensure that the actors in the legal system can make the probabilistic and normative judgments necessary to apply the relevant legal concepts. Moreover, it allows us to more plausibly argue that the law is fair and that it can deter risky behavior (if the reasonable person can anticipate certain errors he or she can perhaps act to prevent them).

For example, products liability law has made manufacturers liable not only for intended use of their products but for foreseeable misuse of their products (American Law Institute, 1998, § 2 R.N.). The idea here is to not let manufacturers dodge liability by claiming they intended for the product to be used in one way even though they should know that people will use it in a different way. Tort law thus makes manufacturers responsible for making judgments about how ordinary people will use their products and taking reasonable precautions to prevent harmful misuses. Similarly, 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 (American Law Institute, 2006, § 2.04 cmt. b).

Both of these doctrines are typically thought to impose a fairly strict standard of liability in which a defendant can be held responsible for accidents that we might say are not their fault. Still, limits exist that track social understandings.

3.2. AI Failure and the "Foreseeability Gap"

The law appreciates 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. Of course, the risk of health complications is not unforeseeable or unpredictable in the ordinary sense of the term. Our conclusion that this outcome is legally unforeseeable reflects a judgment about their probability and the fairness of attributing the harm to driver A's decision to drive. The best solution is to hold no one legally responsible and to fill the gap with insurance.

In some cases, AI errors are unpredictable and unusual in ways that can be both difficult to understand probabilistically and to categorize for the purposes of assigning responsibility. Over time, both issues might be overcome as our understanding of these systems improves and social understandings develop that can inform our judgments of fairness. In the meantime, we believe the application of legal foreseeability to certain AI errors is unclear.

Of course, the law could solve this problem by adopting a high-level view of the legal foreseeability of AI errors. While it is difficult to predict what types of errors will emerge from an AI system it is predictable that they will err. Courts *could* choose to adopt a view of foreseeability that operates at a high level of generality. (Lior, 2019) If so, AI manufacturers will be asked to operate as de facto insurers. But by making no distinction between responsible and irresponsible development and deployment of AI, this approach seems morally and economically flawed.

Leaving the gap open and letting victims bear the losses also seems unsatisfactory. 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, but it's not ideal.

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.

4. How and Why AI Fails

4.1. How AI Errs

We provide a non-exhaustive list of types of failures for AI systems. These failures can occur during training and deployment. Note, we list these types of failures in one section but they have varied liability implications (see Section 5.2 for an example). A model with implicit bias built-in during training is a fundamentally worse product than a model whose performance gradually degrades in an evolving world. Plaintiffs and defendants may point to different foreseeable errors to advance their competing liability arguments.

Learned Model Bias. Modern machine learning techniques require a large amount of data. Some estimate that GPT-4 required over thirteen trillion tokens of input data during training (Walker, 2025). With such large amounts of data during training, the adage garbage in, garbage out remains as strong as ever. These problems exist in a wide-ranging set of fields from criminal justice (Angwin et al., 2016), healthcare (Obermeyer et al., 2019), and insurance outcomes (Huang, 2022). The legal field has devoted significant attention to the topic of model bias, particularly in criminal law (O'Neil, 2017; Mayson, 2018; Páez, 2021; Selbst, 2017). 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. 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, 2018). Judges that use Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a tool for predicting a convicted criminal's recidivism rate, for example, want to ensure that the algorithm uses non-racial reasoning (Angwin et al., 2016). This includes not generating predictions based on variables that are racially coded, such as neighborhood or ZIP code in a highly segregated city. Further complicating the issue, the exact variables used by COMPAS are not known as the model is proprietary (Smith, 2017).

Explainability (Burkart & Huber, 2021) 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, the trend of deeper neural networks with ever-growing parameter counts has made the already difficult task of explainability even harder (Fan et al., 2021). 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 an infraction has occurred, creating a system of retribution rather than avoidance.

Reward Hacking. Models can learn shortcuts or abnormal behaviors to optimize a loss function in unexpected ways. Consider an example from OpenAI of the boat-racing video game CoastRunners (Clark & Amodei, 2016). In the game, the objective is to finish 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 the optimal behavior is 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, one can expect similar failures as AI developers deploy models into the real world.

Out-of-Distribution Inputs. Machine learning models perform incredibly well on withheld testing data on in-distribution samples. However, their success rapidly diminishes on out-of-distribution (OOD) inputs (Yang et al., 2024). It is not surprising that a model trained to differentiate cars from airplanes might fail when provided with a battleship for which it has not assigned an output label. However, the models do not just fail to classify the input as "battleship"—the models often will predict the battleship as either a car or an airplane with high confidence! As models have increased in the number of parameters, the calibration, i.e., the correlation between confidence and accuracy, has often worsened (Guo et al., 2017). OOD inputs often exacerbate this calibration problem (Tomani et al., 2021).

There has been significant research into detecting OOD inputs (Yang et al., 2024). However, the area of research remains active with no solution outperforming all others. Thus, we cannot expect commercial AI systems to adequately mitigate this problem. AI models may and often will, when presented with OOD samples, fail in difficult to detect ways, despite some methods to help agents identify these inputs (Liu et al., 2020; Liang et al., 2018)

Concept Drift. The environment constantly evolves with time and thus so does the distribution of inputs and outputs. The evolution of the input distribution is called concept drift. As the data distribution changes, the model performance degrades. This phenomenon has been discussed in the academic literature in numerous fields such as medicine (Sahiner et al., 2023), cybersecurity (Matejek et al., 2025), and natural language (Koh et al., 2021). This poses a particular problem for AI products purchased without a subscription. These models will almost inevitably degrade in performance as the world changes. Without a subscription service to guarantee model updates, the model will remain stagnant. If the manufacturer of such a model is found liable for an eventual failure, it almost necessitates subscription pricing for every commercially available machine learning tool.

Catastrophic Forgetting. During catastrophic forgetting, a model "forgets" the previous training distribution and struggles to correctly classify examples from that distribution (Chen & Liu, 2018). Ideally, the in-distribution samples of a model is an ever-increasing set but in practice the model's distribution becomes warped. Catastrophic forgetting remains an unsolved problem despite a significant amount of research (Wickramasinghe et al., 2023). Catastrophic forgetting is particularly challenging when a model trained on a given task is updated to perform a novel one (while needing to remain competent in the original task) (Aleixo et al., 2023). This will become an increasing problem as the industry favors base foundational models later tuned for domain-specific tasks (e.g., Archit et al. (2025) fine-tune a segmentation foundation model for electron microscopy images.). Practitioners often find that the re-training updates model weights that were once critical for completing the original task.

Human Alignment Failures. The major artificial intelligence companies today use a technique called reinforcement learning from human feedback (RLHF) to finetune LLMs. 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 numerous 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 (OpenAI, 2023). Perhaps more accurately, companies align models with the values of the creators and their nation states (Alinia AI, 2025; Lu, 2025).

Human alignment is often brittle and even breaks when switching to "low-resource" languages (i.e., languages underrepresented in the training data) (Deng et al., 2024). 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 (Schulhoff). 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 including generating harmful content.

Classification Errors. The most traditional source of error is a classification error where a model fails on an in-distribution input. With the abundance of data, models have, on many problems, approached or surpassed "human accuracy" levels. For example, a meta-analysis suggests that computer models are comparable to dermatologists on identifying melanoma and outperform non-domain clinicians (Salinas et al., 2024). However, any model, just like any expert, may eventually err. There are obscure corner cases, latent causes, and input noise that make these problems inherently difficult. It is unrealistic to assume that an AI model will correctly 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 (Abbott, 2020).

Hallucinations represent another variant of traditional error where a large language model (LLM) will "invent" information in their responses (Kalai et al., 2025). 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 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, for example, but an artifact of the model weights favoring an incorrect statement.

4.2. AI Failures in the Real World

Environmental Context and Autonomous Vehicles. The environmental context can greatly change the liability landscape during AI failure scenarios. Consider the problems that arise from autonomous vehicles and note that, as of the date of this writing, there are no truly fully autonomous cars on the market according to the Society of Automotive Engineers (SAE International, 2021). Consider the following 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 (Krisher, 2023). 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 (Hogan, 2023). A YouTube content creator turns on FSD mode on a windy country road and immediately removes himself to the backseat (see Díaz (2021) for a similar case). 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 will 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 (Tesla, 2026). In the third scenario, the driver is majority at fault for a clear abuse of the self-driving capabilities of the Tesla vehicle.

General versus Specialized Algorithms. Generative adversarial networks (Goodfellow et al., 2014) and diffusion models (Sohl-Dickstein et al., 2015) allow one to create photorealistic images of a variety of topics. 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 (Stability AI). These image generation tools allow one to provide the type (photorealistic, renaissance, baroque, cartoon, etc.) as well as the subject of an image, 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 (Pei et al., 2024). These deepfakes can be used as "evidence" in misinformation campaigns (Department of Homeland Security, Office of Intelligence and Analysis, 2021). In some of the most disgusting abuses of this technology, (ab)users have generated portrayals of famous actors and actresses in pornographic scenes. Some states have added deepfakes to existing revenge pornography laws to provide additional protections for their citizens (Hao, 2021). It seems obvious that tools like Stable Diffusion that specialize in image generation should have internal safeguards from such abuses of their system. Liability for the creation of deepfakes using these tools would 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 they provide to the general public. We refer to these as specialized tools.

These image generating tools are building on existing algorithms in the academic literature. Furthermore, LLMs are trained on a large corpus of academic literature and open source code. One could simply query GPT-5 to produce code that implements a diffusion model. Skeptics will (rightfully) point out that having working neural network code is significantly different than having a frontier model. However, when it comes to deepfakes, one doesn't necessarily need the best generated image software 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 (which is generally seen as a positive for the research community). In these scenarios, who is liable for the creation and proliferation of deepfakes? It would seem that when using a general-purpose tool to construct an algorithm for malicious purposes, the liability shifts to the (ab)user.

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 (Wells, 2023a). Tessa was created by health researchers specifically for this use case and sold to replace the team of human operators previously working the helpline (Wells, 2023b). 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 API to field questions from callers to the helpline. This chatbot similarly fails spectacularly, providing weight loss regiments to those looking for body acceptance. In this 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.

Currently, some companies are focusing on domain-specific AI agents for the professional class, which leads to some unique legal questions. One such example is Harvey, a digital assistant for lawyers (Harvey AI). As an end-user, a lawyer can significantly improve productivity using Harvey when analyzing existing proprietary documents or finding relevant court cases. Despite the intended use case, it is likely that some companies will use Harvey (or equivalents) as their legal counsel. 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 save on legal fees. Furthermore, a company providing such a specialized agent 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).

4.3. Academic Progress and the Foreseeability Gap

Section 4.1 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. Unsurprisingly, the "safeguarding" research trails the cutting-edge products on the market. 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. 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 safeguards. This does not imply that a known problem having an academic sub-discipline requires a fully-implemented solution for every AI system. Rather, the courts operate on an (albeit) fuzzy notion of what precautions a reasonable AI developer would and should employ. The expectation of mitigation falls to the developer as the problems and solutions to these problems become both more well-documented and robust.

5. Overcoming (Un)foreseeability

5.1. A Framework for Foreseeability: Why, not How

To recap, in our view the hard questions of legal liability for AI errors occur where the errors are unpredictable and difficult to assign responsibility for due to the nature of AI systems. Because foreseeability is built into the structure of tort law, existing doctrines struggle to address these hard questions. If fault-based liability is to fill this gap, a modified approach is needed. We propose three general principles for where existing rules run out:

Principle 1: Errors that cannot be traced to a known cause of AI failure should not be the basis for liability. Alternatively stated, traceability is a necessary, but not sufficient condition for liability.

Principle 2: For errors that are traceable, researchers and courts will need to develop appropriate standards for how developers should attempt to prevent harmful errors. And, in some cases, meeting those professional standards should constitute a defense to liability. The relevant professional standard may also vary depending on whether the error is emergent or the result of deliberate design choices.

Principle 3: Where an AI system is deployed by an external party in a way that harms a third party, liability rules should take into account which actor was in the best position to know about the potential for error and to prevent the harm from occurring. In particular, in the case of specialized actors (e.g., doctors) using AI systems sufficient warnings about the potential for and nature of errors should typically suffice to shift responsibility to the specialized actor.

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 speeding and causes an unusual accident. Our proposal would potentially 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.

We think that this proposal balances several important considerations. First, it compensates victims of AI models by limiting the foreseeability gap. Although 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.

We note that these principles leave important details to be worked out. A couple are of particular note. First, when does a category of error become sufficiently recognized to meet our standard? One idea might be to consider professional custom though this may not be a complete solution (Choi, 2023). Second, what level of care suffices to defeat liability raises further questions about professional custom.

5.2. Moving Forward for AI Developers

If an AI developer is going to use unforeseeability as a defense for liability, they need to create systems where the nature of the error can be traced. As discussed in Section 4, many of the so-called unusual errors that AI makes are well studied in the academic literature. Furthermore, these researchers have created some safeguards to protect against these types of failures (e.g., energy-based (Liu et al., 2020) and ODIN (Liang et al., 2018) out-of-distribution detectors can offer some level of protection from such inputs).

Consider the following hypothetical future case: a remote AI agent misdiagnoses a rural patient who otherwise would have made the hours-long trek to the nearest human-operated clinic. The patient later succumbs to his illness, a rare but not unheard of condition for his demographic in that part of the world. His family seeks compensation for his untimely death against the company that produces these roaming "AI clinicians". The manufacturer attests that their AI clinician made a simple classification error. Furthermore, they argue that a reasonably competent "human clinician" in the same specialty could conceivably make the same mistake given the available evidence at the time of examination. During the ensuing court case, the plaintiffs contest the manufacturer's claim by suggesting that the AI agent has implicit bias from the lack of training data from the deceased's racial and socioeconomic demographic. The court now needs to decide, through analyzing the training data composition and inferential decision-making process, whether the error was caused by inherent model bias or a reasonable classification error.

In order for this legal framework to succeed in litigating the above scenario, AI developers need to ensure some level of provenance over their models during both training and deployment. If a developer plans to argue unforeseeability under our modified standard as a defense against an AI error, they need to at least provide the courts (and the plaintiffs) with enough evidence that the error does not trivially fall into existing known failure categories. As discussed in Section 4, errors can occur during training such as the introduction of implicit model bias or through reward hacking. We believe some level of data provenance is required to ensure that lawyers can look back into the training apparatus and determine the source for possible failures like implicit bias. Something like Data Cards (Pushkarna et al., 2022) could enable non-technical experts to gather insights into the types and sources of data.[3] We do not expect the courts to try to retrain any models! However, by looking at the data, one can identify potential causes for failure (e.g., a melanoma detector trained predominantly on lighter skin tones (Montoya et al., 2025)). In the above scenario, the plaintiffs would want to show that even with stronger signals of the illness the AI clinician would fail to diagnose correctly based on the *implicit bias of the model*.

[3]The courts already have existing mechanisms to guarantee the safety of AI developers' trade secrets (Levine & Flowers, 2014).

For provenance during inference, we believe snapshots of model weights and random seeds can help allow future forensic AI experts to recreate the conditions of a failure in the same way that a tire track forensics expert can recreate a car crash to litigate blame. We, note, however, that snapshots and the initial starting conditions may not be enough given the non-associativity of floating point arithmetic (Ingonyama, 2024). Future AI developers should perform the best practices from the PyTorch documentation (PyTorch Contributors) for minimizing non-determinism. Since this is a known phenomena, best practices will continue to evolve and new mitigations may emerge (e.g., see *Software Mechanisms and Recommendations* in Brundage et al. (2020) for a more extensive list of recommendations). In the same way that academic literature may uncover new foreseeable errors (and thus mandate additional AI safety mechanisms), the requirements for provenance may force developers to adjust their engineering practices. The lawyers for the manufacturer company above would try to recreate the output from the AI clinician to identify the *reason(s) for the misdiagnosis*. Drawing on explainable artificial intelligence, they may argue that a reasonable human doctor with relevant domain expertise might have made the same misdiagnosis.

6. Conclusion

AI systems have enormous potential. By "thinking" differently they can solve human problems often better than the typical person, they promise to unlock significant economic value and save lives. But this difference is the source of unpredictable and unusual errors. Where these errors harm others, they produce cases that are hard to fit into traditional legal categories.

That said, we sometimes know why AI fails, even if it is in an unusual way. Therefore, we propose modest adjustments to the legal framework for AI that defaults to traditional tort categories in foreseeable instances with mixed liability in truly novel or unforeseeable instances. Because we require the proof or disproof of the nature of an error, AI developers must provide some level of provenance on their models during both training and inference. Therefore, both plaintiffs and defendants can glean insights into a model's action and make their cases in court as to the cause of failure. As the academic community explores the inner mechanisms of AI agents, we envision a shrinking space of truly unforeseeable errors where traditional tort law can again take over.

LLM Usage Statement. The authors used LLMs to refine some of the text to improve accessibility to both a legal and technical audience. The authors also used LLMs to surface related work for author analysis.

Acknowledgments. Thanks to Dhruv Aggarwal, Zach Fields, Nathaniel Donohue, James Grimmelmann, Patrick Hulin, Peter Kallis, Shelly Simana, Adam Cobb, and Ashish Gehani for helpful conversations and comments on the draft.

References

Abbott, R. *The reasonable robot: Artificial intelligence and the law*. Cambridge University Press, 2020.

Aleixo, E. L., Colonna, J. G., Cristo, M., and Fernandes, E. Catastrophic forgetting in deep learning: A comprehensive taxonomy. *arXiv preprint arXiv:2312.10549*, 2023.

Alinia AI. The alignment question in the age of DeepSeek. https://alinia.ai/the-alignment-question-in-the-age-of-deepseek/, January 2025.

American Law Institute. *Restatement (Third) of Torts: Products Liability*. American Law Institute, Philadelphia, PA, 1998.

American Law Institute. *Restatement (Third) of Agency*. American Law Institute, Philadelphia, PA, 2006.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016.

Archit, A., Freckmann, L., Nair, S., Khalid, N., Hilt, P., Rajashekar, V., Freitag, M., Teuber, C., Spitzner, M., Tapia Contreras, C., et al. Segment anything for microscopy. *Nature methods*, 22(3):579–591, 2025.

Ball, D. W. and Rozenshtein, A. Z. Congress should preempt state AI safety legislation. https://www.lawfaremedia.org/article/congress-should-preempt-state-ai-safety-legislation, June 2024.

Bathaee, Y. The artificial intelligence black box and the failure of intent and causation. *Harv. JL & Tech.*, 31:889, 2017.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., et al. Toward trustworthy ai development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*, 2020.

Burkart, N. and Huber, M. F. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

Chen, Z. and Liu, B. *Lifelong machine learning*. Morgan & Claypool Publishers, 2018.

Choi, B. H. AI malpractice. *DePaul L. Rev.*, 73:301, 2023.

Clark, J. and Amodei, D. Faulty reward functions in the wild. OpenAI. https://openai.com/index/faulty-reward-functions/, December 2016.

Deng, Y., Zhang, W., Pan, S. J., and Bing, L. Multilingual jailbreak challenges in large language models. In *Proceedings of the 12th International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vESNKdEMGp.

Department of Homeland Security, Office of Intelligence and Analysis. Increasing threats of deepfake identities. Technical report, U.S. Department of Homeland Security, 2021. URL https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf. Produced on behalf of the Office of the Director of National Intelligence.

Diamantis, M. E. Employed algorithms: a labor model of corporate liability for AI. *Duke lJ*, 72:797, 2022.

Díaz, J. California man is arrested after riding in back seat of Tesla on autopilot. New York Times. https://www.nytimes.com/2021/05/12/us/california-tesla-backseat-driver.html, May 2021. Accessed: April 21, 2026.

Fan, F.-L., Xiong, J., Li, M., and Wang, G. On interpretability of artificial neural networks: A survey. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 5(6): 741–760, 2021.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Hao, K. Deepfake porn is ruining women's lives. now the law may finally ban it. MIT Technology Review. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/, February 2021.

Harvey AI. Company. https://www.harvey.ai/company. Accessed: April 21, 2026.

Hogan, M. Testing a full self-driving Tesla in Detroit snow looks reckless and embarrassing. Road & Track. https://www.roadandtrack.com/news/a42384858/testing-a-full-self-driving-tesla-in-detroit-snow-looks-reckless-and-embarrassing/, January 2023.

Huang, F. How insurers can mitigate the discrimination risks posed by AI. UNSW BusinessThink. https://www.businessthink.unsw.edu.au/articles/insurance-ai-discrimination, July 2022.

Ingonyama. Solving reproducibility challenges in deep learning and LLMs: Our journey. https://www.ingonyama.com/post/solving-reproducibility-challenges-in-deep-learning-and-llms-our-journey, September 2024.

Kalai, A. T., Nachum, O., Vempala, S. S., and Zhang, E. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Kannegieter, T. Nondeterministic torts: LLM stochasticity and tort liability. *YALE L.J.*, 2026.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Krisher, T. US probes crash involving Tesla that hit student leaving bus. Associated Press. https://apnews.com/article/tesla-school-bus-student-hurt-firetruck-d282a5dd63874f22f5e1a6fc8168801b, April 2023.

Levine, B. L. and Flowers, T. C. Your secrets are safe with us: How prosecutors protect trade secrets during investigation and prosecution. *Am. J. Trial Advoc.*, 38: 461, 2014.

Li, Y. and Goel, S. Making it possible for the auditing of ai: A systematic review of ai audits and ai auditability. *Information Systems Frontiers*, 27(3):1121–1151, 2025.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Lior, A. AI entities as AI agents: Artificial intelligence liability and the AI respondeat superior analogy. *Mitchell Hamline L. Rev.*, 46:1043, 2019.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Lu, D. We tried out DeepSeek. it works well, until we asked it about Tiananmen Square and Taiwan. https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan, January 2025.

Matejek, B., Gehani, A., Bastian, N. D., Clouse, D. J., Kline, B. J., and Jha, S. SAFE-NID: Self-attention with normalizing-flow encodings for network intrusion detection. *Transactions on Machine Learning Research*, 2025.

Mayson, S. G. Bias in, bias out. *YAle lJ*, 128:2218, 2018.

Montoya, L. N., Roberts, J. S., and Hidalgo, B. S. Towards fairness in AI for melanoma detection: systemic review and recommendations. In *Future of information and communication conference*, pp. 320–341. Springer, 2025.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

O'Neil, C. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

OpenAI. GPT-4 system card. Technical report, OpenAI, March 2023. URL https://cdn.openai.com/papers/gpt-4-system-card.pdf.

Páez, A. Negligent algorithmic discrimination. *Law & Contemp. Probs.*, 84:19, 2021.

Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., Zhai, G., Yang, J., and Tao, D. Deepfake generation and detection: A benchmark and survey. *ACM Computing Surveys*, 2024.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 1776–1826, 2022.

PyTorch Contributors. Reproducibility. https://docs.pytorch.org/docs/stable/notes/randomness.html. PyTorch 2.11 documentation; accessed: April 21, 2026.

SAE International. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Standard J3016_202104, SAE International, April 2021. URL https://www.sae.org/standards/j3016_202104-taxonomy-definitions-terms-related-driving-automation-systems-road-motor-vehicles.

Sahiner, B., Chen, W., Samala, R. K., and Petrick, N. Data drift in medical machine learning: implications and potential remedies. *The British Journal of Radiology*, 96 (1150):20220878, 2023.

Salinas, M. P., Sepúlveda, J., Hidalgo, L., Peirano, D., Morel, M., Uribe, P., Rotemberg, V., Briones, J., Mery, D., and Navarrete-Dechent, C. A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. *NPJ digital medicine*, 7(1):125, 2024.

Schneier, B. and Sanders, N. E. AI mistakes are very different from human mistakes. IEEE Spectrum. https://spectrum.ieee.org/ai-mistakes-schneier, January 2025.

Schulhoff, S. DAN (do anything now). Learn Prompting, https://learnprompting.org/docs/prompt_hacking/offensive_measures/dan.

Selbst, A. D. Disparate impact in big data policing. *Ga. L. Rev.*, 52:109, 2017.

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., and Kenton, Z. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022.

Sharma, C. AI's hippocratic oath. *Wash. UL Rev.*, 102:1101, 2024.

Smith, R. A. Opening the lid on criminal sentencing software. Duke Today. https://today.duke.edu/2017/07/opening-lid-criminal-sentencing-software, July 2017.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Stability AI. Stable image. https://stability.ai/stable-image. Accessed: April 21, 2026.

Tesla. *Model 3 Owner's Manual*. Tesla, 2026. URL https://www.tesla.com/ownersmanual/model3/en_us/. Model 3 (2024+), software version 2026.8; accessed: April 21, 2026.

Tomani, C., Gruber, S., Erdem, M. E., Cremers, D., and Buettner, F. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10132, 2021.

Vladeck, D. C. Machines without principals: liability rules and artificial intelligence. *Washington Law Review*, 89: 117, 2014.

Walker, II, S. M. Everything we know about GPT-4. KLU. https://klu.ai/blog/gpt-4-llm, September 2025.

Warren, Z. GenAI hallucinations are still pervasive in legal filings, but better lawyering is the cure. https://www.thomsonreuters.com/en-us/posts/technology/genai-hallucinations/, August 2025.

Weil, G. Tort law should be the centerpiece of AI governance. Lawfare, https://www.lawfaremedia.org/article/tort-law-should-be-the-centerpiece-of-ai-governance, August 2024.

Wells, K. An eating disorders chatbot offered dieting advice, raising fears about AI in health. NPR Health Shots. https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea, June 2023a.

Wells, K. National eating disorders association phases out human helpline, pivots to chatbot. NPR Health Shots. https://www.npr.org/sections/health-shots/2023/05/31/1179244569/national-eating-disorders-association-phases-out-human-helpline-pivots-to-chatbo, May 2023b.

Wickramasinghe, B., Saha, G., and Roy, K. Continual learning: A review of techniques, challenges, and future directions. *IEEE Transactions on Artificial Intelligence*, 5(6):2526–2546, 2023.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

Yoshikawa, J. Sharing the costs of artificial intelligence: Universal no-fault social insurance for personal injuries. *Vand. J. Ent. & Tech. L.*, 21:1155, 2018.