THE FORESEEABILITY PARADOX

Sunayana Rane

Abstract

Technical AI safety researchers, regulators, and policymakers have a common goal: to protect us from the harms arising from increasingly ubiquitous AI systems in the real world. For technical researchers, this often entails meticulously studying the behavior of existing AI systems and building better, safer ones. For regulators, policymakers, and courts, it involves effectively regulating and adjudicating this space of breakneck technological development. And yet a paradoxical challenge makes this nearly impossible: humans cannot instinctively predict harmful AI behavior, and the parties who could, with appropriate time and resources, help foresee harmful AI behavior are the ones currently least incentivized to do so. Much of the danger in AI behavior is that it is often dissimilar from human behavior in unexpected ways. Drawing together empirical evidence from AI interpretability and cognitive science, this work explains why humans should not be expected to foresee harmful AI actions. Instead, the foreseeability of AI harms largely depends on the due diligence of the AI's creators – due diligence they are currently disincentivized from undertaking. Fortunately, the legal concept of foreseeability also gives us the foundation for the solution: responsibility for preventing or recompensing and AI harm should center on whether, and to what extent, each actor behaved responsibly with their knowledge of the tendencies of the AI system in question.

TABLE OF CONTENTS

INTRODUCTION

AI regulation efforts and litigation can seem like wading through quicksand – just as courts and legislators are forming a legal consensus about a particular aspect of the technology, the technology changes beneath our feet. The rapid pace of the technological development and the scattered, varied attempts at reining it in with the rule of law beg the question: are we fighting a losing battle?[1] The goal of this article is to analytically illustrate why it doesn't have to be this way: By unifying strategies to solve the foreseeability paradox, we can build technologically-informed legal paradigms to effectively regulate and adjudicate AI harms.

Unlike the rapidly-multiplying state laws and regulations on AI,[2] many of which interact with already complex and outdated product liability regimes[3] to create a legal quagmire distanced from the technical realities of AI, foreseeability is an abstract concept – it is not attached to one type of law, nor is it exclusively in the domain of a particular regulatory agency. It generously assists in doctrinal legal reasoning in tort law, criminal law, contract law, and administrative law. When engaging with a rapidly-evolving technological space, this very quality of abstractness gives it a powerful flexibility. By addressing the foreseeability paradox, researchers and practitioners alike can use the legal concept of foreseeability as a starting point for a bidirectional legal and technical effort towards safe, responsible, and human-compatible AI systems.

Most technical AI safety research is well positioned to feed directly into a foreseeability-centered legal analysis. The technical AI safety research community ranks understanding AI behavior among its key goals.

---

[1] Wissam Salhab, et al. "A systematic literature review on AI safety: Identifying trends, challenges and future directions." *IEEE Access* (2024).

[2] *See, e.g.,* California's SB53 (2025), Oregon's House Bill 2748 (HB 2748) (2025) and New Jersey SR52 (2026) (recent state laws or regulations governing an incomplete slice of AI use)

[3] *See, e.g., Benavides v. Tesla, Inc.*, No. 21-cv-21940, 2025 WL 1768469 (S.D. Fla. Jun. 26, 2025) (self-driving AI wrongful death lawsuit brought under Florida's product liability law)

Technical research aims to characterize, study, and understand human responses to AI – in order to steer AI development towards safer and more human-compatible behaviors. In practice, the information produced from such technical advancements answer questions like: Would AI creators have known that this self-driving AI would run over a pedestrian, and how could they have prevented it? Were state-of-the-art evaluations carried out to improve the foreseeability of potential harms? Did the AI creators conduct the proper due diligence to prevent foreseeable harms, or were they willfully blind? What kind of technical research can help increase the foreseeability of AI harms, and how can we incentivize it?

When it comes to AI creators absolving themselves of responsibility of their AI system causing harm, technical research is increasingly helping us answer the question, "Did they know? Should they have known?" Courts and regulators can incentivize and benefit from the rapid progress of such technical research by centering their analyses and regulations on foreseeability.

A common move in the playbook for defending against an allegation of AI harm is to argue that the human interacting with the AI should have prevented the harm. The defense's argument in the recent Tesla Autopilot litigation provides a clear example of this: Tesla argued that the human driver should have been paying attention and stepped in to prevent the self-driving AI's fatal mistake.[4] The jury found this blame-shifting unconvincing, reflecting an emerging pattern in the fight between AI creators and consumers regarding who is responsible for the harms caused by unpredictable AI behavior.

Parts I-III of this article draw together technical AI research, cognitive science, and foundational principles underlying doctrinal foreseeability analyses to explain the technical nature of the foreseeability paradox: why humans should not be expected to foresee harmful AI actions, and why the foreseeability of AI harms largely depends on the due diligence of the AI's creators.

Fortunately, the legal concept of foreseeability also gives us the foundation for the solution: the legal responsibility for preventing or recompensing and AI harm should center on whether, and to what extent, each actor behaved responsibly with their foreseeable knowledge of the tendencies of the AI system involved. The remainder of this article uses the legal concept of foreseeability as a cornerstone upon which to build a dual

---

[4] *Id.* (stating defense argument that human should have intervened to prevent Tesla Autopilot crash)

framework in which regulators/legislators and courts both have a role to play.

Drawing once again from empirical research, Part IV of this article structures the role of regulators – state and federal legislators and administrative agencies – to force information transparency regarding AI behaviors. This includes requiring technologically-informed tests, evaluations, and simulations before an AI system is put in contact with consumers. It also includes requirements for transparency about the kinds of AI evaluations carried out by the AI creator and by authorized third-parties, and the data resulting from those evaluations. Forcing this information for public disclosure prevents the willful blindness and harmful secrecy that can limit the societal foreseeability of harmful behaviors resulting from the AI's widespread use.

Not all harms are preventable, and Part IV of this article also examines the role of courts in adjudicating disputes regarding AI harms that do come to pass. Regulators can use foreseeability as a guiding principle to force transparency and scientific best practices for AI safety. Courts will then be empowered to use this information to determine legal responsibility based on each actor's foreseeability of the harms, and opportunities to prevent foreseeable harms. Parts II and III are closely tied to the scientific understanding of humans and AI systems that we now have, as illustrated in Part I. The flexibility of a foreseeability-based approach extends not only to disciplines within the law, but also to a variety of real-world contexts – this framework is designed to be equally effective for self-driving cars and racially-biased hiring algorithms, for AI-driven teen suicides and for healthcare AI misdiagnoses.

Building on the legal concept of foreseeability will help courts and legislators anticipate, and assign responsibility, for past and future AI harms. As technical AI researchers fill in the details about what AI creators and ordinary consumers can reasonably foresee about AI harms, a legal strategy centered on foreseeability will be able to flexibly incentivize and adapt to this new information about AI behavior. In the process, by building a legal framework that embraces the technical nuances of the foreseeability paradox when it comes to sophisticated and unpredictable AI behavior, regulators and courts can build an improved doctrine that will stand the test of future years of AI development.

I. THE ORIGINS OF FORESEEABILITY

Foreseeability's load-bearing role in the construction of tort, contract, and criminal liability reveals a fundamental underlying principle we accept to be true in our societal interactions with one another: acting responsibly

entails necessarily considering the consequences of our own actions and of others' reactions to them.

Foreseeability and "expectability" emerged and crystalized as common-law concepts – that is, they emerged over time as necessary components of the analysis judges would make in absence of any preexisting written legal code.[5] As a legal scholar observed nearly a century ago, the emergence of foreseeability meant that the "whole idea of risk or threat is comprehended in the notion of foresight—foresight in the sense of the probability of harm resulting from conduct."[6] Foreseeability introduced, for the first time, a concrete expectation that we mentally simulate the chain of cause-and-effect resulting from any action we choose to take, including the expected reactions of others.

Foreseeability emerged as a legal formalism because the realities of human behavior suggested that it should. We would not, for example, wish to punish one another for entirely unforeseeable causal consequences of our actions. However, we would also wish to impose an affirmative duty for everyone to consider the probable cause and effect of our actions. But what about the realities of AI behavior, sometimes only revealed through details of their technical implementation? How can we adapt the principles underlying the origins of the legal concept of foreseeability to a world where AI systems increasingly encroach into tasks and decisions hitherto relegated solely to human decision-makers?[7] How can we adapt a legal concept designed with a common-sense understanding of human behavior to the new and uncertain realm of AI behavior? These are the key questions this work considers.

---

[5] Fowler Vincent Harper. "Foreseeability factor in the law of torts". In: Notre Dame Law. 7 (1931), p. 468. (explaining the common law origins of foreseeability).

[6] *Id.*

[7] *See, e.g., Danielle Keats Citron and Frank Pasquale. "The scored society: Due process for automated predictions." Wash. L. Rev. 89 (2014): 1. (explaining that automated tools are slowly taking over everything from insurance assessment to healthcare decisions, from mortgage decisions to financial planning)*

I. FORESEEABILITY OF AI BEHAVIOR

The reasons for an AI system's decisions, actions, and outputs are usually not obvious even to technical researchers,[8, 9] and some would even argue that there aren't "reasons" at all in the sense of conscious, causal reasoning.[10, 11] Modern AI is incredibly powerful because it has billions of moving pieces inside its "brain," and although these AI systems are used without much restraint, these complex internals are incredibly poorly understood. For example, if a self-driving car crashes into a pedestrian, it is unlikely that even the AI's creators can know exactly what went wrong and caused the crash. Rather, the causal chain leading to AI decisions is a confluence of model inputs, training data and its effects on learned internal weights and biases, constraints produced by architecture, objective function, and training methodology, and any safety and bias checks (or lack thereof).[12] And these factors are often quite removed from the expectations of common sense and good judgment that we have of one another in society.[13]

---

[8] *See, e.g.,* Hannah Rashkin et al. *Measuring attribution in natural language generation models.* COMPUTATIONAL LINGUISTICS 49.4 (2023). (explaining why even the most technical folks don't understand the reasons behind an AI's decisions)

[9] Eliana Pastor et al. *Beyond Input Attribution: A Hands-On Tutorial to Concept-Based Explainable AI and Mechanistic Interpretability*. PROCEEDINGS OF THE 31ST ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING V. 2. (2025). (one of many technical research papers attempting to understand the internal workings of AI models)

[10] Jaan Aru, Matthew E. Larkum, and James M. Shine. "The feasibility of artificial consciousness through the lens of neuroscience." Trends in neurosciences 46.12 (2023): 1008-1017. (Concluding that "while fascinating and alluring, LLMs are not conscious and will likely not be conscious soon").

[11] *See also* Melanie Mitchell's extensive work on empirical documenting how AI models don't "reason" in a human-like sense, including: Melanie Mitchell. *Artificial intelligence: A guide for thinking humans*. Penguin UK, 2019, and Melanie Mitchell and David C. Krakauer. "The debate over understanding in AI's large language models." *Proceedings of the National Academy of Sciences* 120.13 (2023): e2215907120.

[12] Ian Goodfellow, et al. *Deep Learning*. Vol. 1. No. 2. Cambridge: MIT Press, 2016. (canonical textbook on modern AI methods, which lays out the various components, training paradigms, and other messy, opaque factors that contribute to an AI model's final behavior)

[13] Sunayana Rane. "The Reasonable Person Standard for AI." *Forty-first International Conference on Machine Learning*. 2024. (documenting the ways in which AI behavior differs from the reasonableness standards we frequently apply to humans).

Harper's 1931 treatise on the rise of foreseeability as a legal concept goes on to explain how its invention reflects this shared societal expectation of reasonably cautious, responsible, common-sensical behavior:

> "Experience suggests the danger incident to certain activity. Not particular experience of particular individuals, but general experience—experience that often defies analysis—the multitude of factors, knowledge, hunches, instincts or what they may be called, the common sense that makes social intercourse possible, all operate to prompt the 'ordinary reasonable man' that harms are 'probable' or 'natural' as normal results of certain situations and certain conduct."

Not only do today's AI systems lack common-sense,[14] they are dangerously well-positioned (designed, even) to persuasively mimic it. The flaws are often difficult to see until it is too late -- that is, unless you are looking for them. That is why it is so important that today's legal formalisms incorporate the realities of not only human behavior, but also AI behavior.

Foundation models (such as large language models, or LLMs) often obfuscate, hallucinate, and provide inaccurate information.[15] When asked about their own reasons for their decisions, LLMs can provide persuasive-sounding but entirely unfaithful responses.[16] One might counter that humans do this too -- we lie about our real intentions and motives, and empirical evidence suggests that we take actions instinctively, before our rational decision-making centers explain why we are doing what we are doing (even prompting pop science headlines stating that we lack free will entirely)[17]. However, to one another, humans are easier to read than an AI system is -- we share biology, a certain foundational set of experiences

---

[14] Ronald J Brachman and Hector J Levesque. Machines like us: toward AI with common sense. MIT Press, 2023. (explaining why machines don't currently have human-like common sense)

[15] Vipula Rawte, Amit Sheth, and Amitava Das. "A survey of hallucination in large foundation models". In: arXiv preprint arXiv:2309.05922 (2023) (overview of hallucination tendencies of today's most sophisticated AI models, called "foundation" models)

[16] Miles Turpin et al. "Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting". In: Advances in Neural Information Processing Systems 36 (2024) (recent work showing that models cannot reliably explain themselves)

[17] Benjamin Libet. "Do we have free will?" In: Journal of consciousness studies 6.8-9 (1999), pp. 47–57.

(eating, laughing), and often substantial culture (going to school, sitting down to dinner, wearing clothes), in addition to an evolutionary social acuity, all of which allows us to make reasonable inferences about one another's true intentions and expected behavior (a capacity known as *theory of mind*)[18]. It is perhaps our very acuity for theory of mind inferences about fellow humans that makes us particularly susceptible to anthropomorphizing AI systems, and believing them when they sound like a believable human[19, 20, 21, 22] —something they are explicitly trained to do.

Existing legal frameworks for causality assign blame and responsibility for human behavior in a way that does not easily or smoothly translate to AI behavior. Our notions of foreseeability, reasonableness, and causal chains for human behavior need to be adapted to the realities of AI training and capabilities. At the same time, these legal formalisms also provide a useful starting point for how we should think about causality in AI behavior. Empirical technical research which investigates the causal chains of AI behavior, including through interpretability research, chain-of-thought reasoning, and behavioral studies inspired by analogous cognitive studies of human behavior, can help us build on the foundation provided by these existing legal formalisms.

The AI systems we focus on in this analysis include foundation models, frontier models, LLMs -- in other words, sophisticated models that can truly be said to exhibit "behaviors" instead of just a narrow set of outputs for one particular task. Some of the principles we discuss, however, do apply to simpler AI models as well, and can provide useful guidance in any situation where we find AI (simple or complex) substantially replacing human decision-making in the real world.

---

[18] Chris Frith and Uta Frith. "Theory of mind". In: Current biology 15.17 (2005), R644–R645 (background on the cognitive science principles of theory of mind)

[19] Nicholas Epley, Adam Waytz, and John T. Cacioppo. "On seeing human: a three-factor theory of anthropomorphism." *Psychological review* 114.4 (2007): 864. (expounding one recent theory for why humans, in our instinct to apply theory of mind to understand other humans, also read human-like intentions, beliefs, and goals where there are in fact none)

[20] Linnda R. Caporael, and Cecilia M. Heyes. "Why anthropomorphize? Folk psychology and other stories." *Anthropomorphism, anecdotes, and animals* (1997): 59-73. (explaining how and why humans anthropomorphize).

[21] Nicholas Epley et al. "When we need a human: Motivational determinants of anthropomorphism." *Social cognition* 26.2 (2008): 143-155. (explaining the potentially dangerous human tendency to anthromoporphize when we need human-like emotions and thoughts).

[22]Indrit Troshani, et al. "Do we trust in AI? Role of anthropomorphism and intelligence." *Journal of Computer Information Systems* 61.5 (2021): 481-491. (documenting when and how humans tend to trust AI, and the role of anthromophorism in the process of eliciting trust)

II. FORESEEABILITY AND RESPONSIBILITY

Foreseeability's most significant role is perhaps in tort law. Inherent in the legal principles of proximate and but-for causation is the assumption that to cause an injury is to be responsible for it.[23,24,25,26] The legal concept of foreseeability narrows this field of responsibility.[27,28]

Foreseeability restricts the responsibility associated with an injury to only those harms which were reasonably foreseeable to an ordinary person; we don't want to punish or deter for harmful outcomes that are out of a person's control. Herein lies the first problem: while foreseeability makes us responsible for foreseeable actions, when it comes to AI behavior, many actions are not foreseeable in the human sense.[29] In humans, the foreseeability requirement depends on an ability to mentally "time-travel"/step forward a causal chain[30] and on the assumption that humans have an inherent inclination and responsibility to conduct this mental time travel to consider the consequences of their actions. This ability is understood to form the basis of human foresight,[31] the very capacity to

---

[23] Carolina Sartorio. "Causation and responsibility." Philosophy Compass 2.5 (2007): 749-765. (one of several works tying causation to the fundamental principle of responsibility for a harm)

[24] Michael S. Moore. "Causation and responsibility." *Social Philosophy and Policy* 16.2 (1999): 1-51. (one of several works tying causation to the fundamental principle of responsibility for a harm)

[25] Shelly Kagan. "Causation and responsibility." *American Philosophical Quarterly* 25.4 (1988): 293-302. (work applying legal analysis to tie causation to the fundamental principle of responsibility for a harm)

[26] Alex Kaiserman. "'More of a cause': Recent work on degrees of causation and responsibility." *Philosophy Compass* 13.7 (2018): e12498. (more recent empirical work closely analyzing how people think about causation and responsibility, finding that one must have a role in causing the harm to be responsible for it)

[27] Benjamin C Zipursky. "Foreseeability in breach, duty, and proximate cause". In: Wake Forest L. Rev. 44 (2009), p. 1247. (presents a view of foreseeability as a narrowing construct)

[28] W Jonathan Cardi. "Reconstructing foreseeability". In: BCL Rev. 46 (2004), p. 921. (presents a view of foreseeability as a narrowing construct)

[29] Gesa Wiegand et al. "'I'd like an explanation for that!" Exploring reactions to unexpected autonomous driving." *22nd International Conference on Human-computer Interaction with Mobile Devices and Services*. 2020. (empirical study documenting showing that humans are often taken aback and want explanations when faced with unexpected, unpredictable behaviors of self-driving cars)

[30] Thomas Suddendorf and Michael C Corballis. "The evolution of foresight: What is mental time travel, and is it unique to humans?" In: Behavioral and Brain Sciences 30.3 (2007), 658 pp. 299–313. (explaining how humans mentally time travel as a key aspect of their cognitive function)

[31] *Id.*

which Harper attributed our probabilistic reasoning and dependence on foreseeability as a legal concept nearly a century ago.

Currently, there is ample evidence that AI's causal chain (or even the uninterpretable concepts that form the links in a causal chain)[32] do not match our own, but rather demonstrate spurious correlations with training data or unfortunate consequences of objective functions used for training.[33] For example, an AI system learned to classify images of wolves remarkably well; but upon closer inspection, researchers discovered that it was really learning to positively classify images with a snowy background, which were highly correlated with wolf photos in the training data.[34] AI researchers summarized difference between a human understanding of the underlying concepts and the AI's deceptively successful-looking learned shortcut: "a wolf is still a wolf even when it is in Grandmother's house."[35]

This is only the beginning of the problem. Not only does it take substantial technical work to even attempt to understand the true causal chain underlying AI behavior (the "why" of an AI's decision or action)[36], but in modern LLMs, the model's own answers are often deceptively persuasive but entirely inaccurate when describing the model's inner workings.[37] Taken together, these technical realities highlight the need for a legal framework for foreseeability of AI decision-making, one that intelligently incorporates technical requirements for model transparency to the extent possible, and rethinks what should be considered "reasonably" foreseeable for AI behavior.

---

[32] Finale Doshi-Velez and Been Kim. "Considerations for evaluation and generalization in interpretable machine learning." *Explainable and interpretable models in computer vision and machine learning*. Springer International Publishing, 2018. 3-17. (documenting the well-understood reality that the internals of AI models are not human-interpretable, and laying out the technical research that is trying to close this gap)

[33] *Id.*

[34] Drew Roselli, Jeanna Matthews, and Nisha Talagala. "Managing bias in AI". In: Companion proceedings of the 2019 world wide web conference. 2019, pp. 539–544. (providing the examples of spurious correlations leading to bias in AI)

[35] *Id.*

[36] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning." *arXiv preprint arXiv:1702.08608* (2017). (laying out the challenges and paths forward towards technical AI interpretability)

[37] Miles Turpin et al. "Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting". In: Advances in Neural Information Processing Systems 36 (2024) (recent work showing that models do not reliably explain themselves and instead provide inconsistent explanations of their internal workings)

A. The Legal Realist View of Foreseeability and Proximate Cause

Legal realists have long argued that foreseeability and causation are both legal concepts created to get at the underlying *responsibility* for a harm. This school of thought suggests that the formalisms around proximate cause and foreseeability are merely post-hoc rationalizations of a court's implicit moral judgements.[38] However, while many legal theorists have often viewed this as a shortcoming, it can also be seen as the reverse: an example of the law developing to incorporate the realities of a situation and still deliver justice, and later constructing formalisms to embody the letter and spirit of the caselaw. Legal realists have criticized the entire doctrine of proximate cause as simply used "to disguise the moral judgment made by the judge" regarding the "responsible cause" of the harm in question.[39] However, the creation of a legal concept as a post-hoc explanation of responsibility is not necessarily a bad thing -- it may, in fact, be a useful blueprint for considering responsibility in AI behavior. In the formative era for the doctrines of modern tort and criminal liability, foreseeability emerged as an important and recurring part of the analysis. This served as an indication that it was relevant and necessary to the analysis.

We are once again in such a formative time, if not a *more* formative time, with the advent of AI that exhibits human-level, and often human-like behavior. We should consider how to replicate this emergence at an equally fundamental level of legal formalism for murkier causal chains underlying AI behavior. These AI systems have already taken over responsibilities from human counterparts in a multitude of domains from healthcare to defense, finance to art, social work to agriculture. As they are trusted with responsibilities previously relegated only to humans, analyzing their actions (and the consequences of their actions) becomes increasingly complex. The existing legal frameworks around foreseeability and causation of human actions are woefully insufficient for the technical realities of AI systems. We need new legal theories and "empirical jurisprudence"[40] designed specifically for AI systems.

---

[38] Joshua Knobe and Scott Shapiro. "Proximate cause explained". In: The University of Chicago Law Review 88.1 (2021), pp. 165–236. (legal realist view of why proximate cause emerged as a legal concept)

[39] *Id.*

[40] *Id.*

B. Foreseeability and causation in the law

Foreseeability was codified as a legal concept as far back as *Donoghue v. Stevenson*,[41] and from there further back to its philosophical origins. As Helen Scott writes in *The History of Foreseeability*, "in the context of Roman *culpa* foreseeability functioned as a technique for determining the avoidability of the harm—essentially a causal inquiry."[42] The notion of considering the consequences of our actions is prevalent in tort law, where the legal concept of foreseeability is most widely discussed. It is important to note that the formal legal notion of foreseeability is different from the common meaning of the word, and even from the meaning that is often applied in the cognitive science literature when alluding to humans foreseeing one another's actions. There is, however, an underlying entanglement with causal chains of events in both conceptualizations.

There are four elements of an unintentional tort: duty of care, breach of this duty, causation of the harm, and damages inflicted by the harm. The formal legal concept of foreseeability frequently comes up in the third element: causation. For a defendant to be liable for a harm, their actions must be the proximate cause of the harm, and the harm must also be foreseeable. Foreseeability is therefore traditionally tied to the analysis surrounding the proximate cause of a harm.

Although foreseeability itself narrows the scope of potential liability, cases like *Tarasoff v. Regents of the University of Californi*a}[43] further expanded the potential reach of foreseeability by attaching it not only to causation but to the *duty* component of a tort. In *Tarasoff*, psychiatric medical providers were held liable for permitting the release of a patient who threatened to murder a particular victim upon release -- he was released anyway, and subsequently murdered her. *Tarasoff* held, and many courts still maintain, that the extremely foreseeable nature of the harm contributed to creating a special duty of the psychiatrists to prevent the harm. Foreseeability, and the causal chains we expect one another to reason through before taking considered action, suddenly became a part of the duty of care we owe to one another. Indeed, in the years since *Tarasoff*, legal scholars have attempted to either reconcile or disavow the increasing

---

[41] Donoghue v. Stevenson [1932] AC 562. UKHL 100, SC (HL) 31, AC 562, All ER Rep 1. 1932. (referred to as the origin of foreseeability as it appears as a legal construct in our modern legal system, although the real origins of the idea are older)

[42] Helen Scott. "The History of Foreseeability". In: Current Legal Problems 72.1 (2019), pp. 287– 648 314. (presenting foreseeability as a causal inquiry)

[43] Tarasoff v. Regents of the University of California. 17 Cal. 3d 425, 551 P.2d 334, 131 Cal. 660 Rptr. 14. 1976. (case in which foreseeability featured prominently, as explained in main text)

importance of foreseeability in determining duty.[44] Once again, however, the intuitions hold: An expectation of preventing foreseeable harms implies that we are all expected to understand, anticipate, and *consider* the consequences of actions. In other words, we all have a *duty* to do a certain amount of *causal* reasoning about the world. Unfortunately, the foreseeability paradox of AI harms ensures that the ordinary consumer (or independent public bodies) do not have the crucial information needed to make a reasonably accurate judgment regarding the foreseeability of AI harms.

C. How society thinks about foreseeability and responsibility

Quite apart from the legal concept of foreseeability as a term of art, advances in cognitive science have given us a window into the word in its common meaning in society. This common meaning is important to consider because it reflects the views of a society which is now on the receiving end of AI behavior, and it will shape how this society views the foreseeability of that AI behavior.

For example, studies in cognition have found that we tend to rate intentional actions as both more causal and more blameworthy than unintentional ones. We also tend to rate highly foreseeable actions as more causal and more blameworthy, and to rate an outcome as more foreseeable if the harm was caused to a vulnerable party, for example to a child or a disabled person.[45] The evidence shows that we are susceptible both to bias and to flexibility in our view of the causal chain leading to certain harms.

Furthermore, a growing body of cognitive science research has shown that our notions of causality are norm-dependent: "a norm-violating agent is seen as more causal than its non-norm-violating counterpart."[46]. This is known as the Norm Effect. Perhaps there is something in the inherent surprise (the lessened foreseeability, if you will) that results from the violation of the norm that affects our causal judgement regarding the outcomes. This tendency could also introduce bias in our judgement of the actions of others.

---

[44] W Jonathan Cardi. "Reconstructing foreseeability". In: BCL Rev. 46 (2004), p. 921. (providing a literature review of scholars reconciling or disavowing foreseeability as a factor in determining duty)

[45] David A Lagnado and Shelley Channon. "Judgments of cause and blame: The effects of intentionality and foreseeability". In: Cognition 108.3 (2008), pp. 754–770. (empirical cognitive science research showing how humans interpret the foreseeability of one another's actions)

[46] Levin Güver and Markus Kneer. "Causation, foreseeability, and norms". In: Proceedings of the 45th annual meeting of the cognitive science society. 2023, pp. 888–895. (empirical cognitive science research about human foreseeability)

Although their objections predated these scientific studies, the results would seem to add credence to the legal realists' view that causal judgements are designed simply to mask moral judgements. At the very least, legal scholarship that has considered these developments has come to the conclusion that causal judgements and moral judgements should be considered together when analyzing how the formalisms surrounding foreseeability and proximate cause arose.

Thus, both empirical evidence and legal scholarship have supported the idea that causal judgements underlying a verdict are quite entangled with moral judgements; it follows that this could be, and perhaps should be, the case with AI.

Finally, regarding human cognitive biases when interacting with AI in general, there is once again ample evidence to show that we instinctively anthropomorphize AI systems and credit their words and actions with human-like meaning – often, undeservedly. When interacting with fellow humans, we exercise theory of mind – we model one another's mental states, and we understand that other people have their own thoughts, feelings, motivations, goals, and experiences.[47] Theory of mind has canonically been seen as a sign of particular intelligence in humans and a handful of other species. Unfortunately, this very tendency to automatically have theory of mind is our undoing when it comes to interacting with AI. We tend to believe the AI system that sounds linguistically human-like (which it is trained to do) *is,* in fact, human-like in terms of having similar thoughts, motivations, and indeed a similar causal understanding of the world that we take for granted in fellow humans.

This human tendency far predates the advent of modern language models like ChatGPT—indeed it has been recorded as far back as the simplistic chatbot ELIZA from the 1960s, designed to sound like a trained psychotherapist.[48][49] Originally designed to be a demonstration of how simple modern AI truly was at the time, the system had the opposite effect -- people began trusting ELIZA with their innermost thoughts and feelings. A modern take on ELIZA, recently released and running a far more sophisticated deep learning-based language model of the ilk of today's

---

[47] Chris Frith and Uta Frith. "Theory of mind". In: Current biology 15.17 (2005), R644–R645 (background on the cognitive science principles of theory of mind)

[48] Joseph Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine". In: Communications of the ACM 9.1 (1966), pp. 36– 672 45. (the original technical paper describing ELIZA)

[49] Jeff Shrager. "ELIZA Reinterpreted: The world's first chatbot was not intended as a chatbot at all". In: arXiv preprint arXiv:2406.17650 (2024) (explaining how ELIZA was intended to show shortcomings of AI systems at the time)

LLMs, is alleged to have contributed to the suicide of an interlocutor, helping him believe he should end his life to stop contributing to the climate crisis.[50]

Empirical studies of both humans and AI systems have begun to document the shifts in human behavior, responses, and perceptions when AI "misbehaves," as it inevitably does. AI-driven decision-making has already permeated decisions of whether someone gets a mortgage[51], whether they get state-allocated aid due to medical disabilities[52], and whether they survive a self-driving car crash.[53] These empirical observations go to the very heart of foreseeability, and its foundational principle, responsibility. One empirical study found that in human-robot teams, when a robot makes an unexpected move or error, the human perceives the robot as having a greater attribution of responsibility than themselves.[54] In other words, "[i]n a situation where a robot teammate does something unexpected […], it is likely that the perceived accountability […] of the robot will increase and the accountability of the human teammate will decrease."[55] Contrary to how we would think about a fellow human, we do not take responsibility for having to anticipate the AI's behavior.[56]

Perhaps there is wisdom to this human instinct. Our societal and legal frameworks for foreseeability and responsibility already have a baked-in assumption that humans model and predict other humans' behavior. We are certainly not perfect at this, but we do expect it of ourselves, and we have the biological and social capacity to do so (in fact, the literature on humans

---

[50] Euronews. "Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to 598 stop climate change." https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop600 climate-. (news article explaining the suicide and the AI's alleged role in it)

[51] Aaryan Gupta,, et al. "AI in Financial Decision-Making." International Conference on Computer Vision and Robotics. Singapore: Springer Nature Singapore, 2024. (explaining AI's increasing role in life-changing financial decisions such as lending credit)

[52] *K.W. v. Armstrong*, 789 F.3d 962 (9th Cir. 2015). (case in which Idaho ACLU successfully sued against AI decision-making denying medical aid to a cerebral palsy patient)

[53] *Benavides v. Tesla, Inc.*, No. 21-cv-21940, 2025 WL 1768469 (S.D. Fla. Jun. 26, 2025)

[54] Joseph B. Lyons, Izz Aldin Hamdan, and Thy Q. Vo. "Explanations and trust: What happens to trust when a robot partner does something unexpected?." *Computers in Human Behavior* 138 (2023): 107473. (empirical study showing human perception of AI responsibility for unexpected behavior)

[55] *Id.*

[56] Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. "Psychological roadblocks to the adoption of self-driving vehicles." *Nature Human Behaviour* 1.10 (2017): 694-696. (empirical study showing that humans blame self-driving cars for mistakes more than they do fellow humans)

anthropomorphizing non-human entities shows that we can *help* modeling theory of mind). However, this instinct does not extend to AI models, and for good reason – these systems are unpredictable in ways that humans are not. A human put in charge of a vending machine would not fill it with inedible metal tungsten cubes, as Claude, Anthropic's most advanced LLM, recently did.[57] This kind of incident, far from a rarity, simply illustrates an underlying technical reality of unpredictable AI behavior that is far from our human norms. It would be unreasonable, and impossible, to ask humans to foresee this behavior, and bake that foresight into their actions, in the same way that we instinctively do for other humans.

We are only beginning to understand the realities of our mismatched sense of human-like meaning in interactions with AI and the true nature of these systems which are designed to *sound* reasonable. Our increasingly deeper understanding of what assumptions we make about causality -- our own causal understanding of the world, and the one we (mistakenly) attribute to AI systems -- should guide the advent of legal frameworks of responsibility for AI systems' actions.

D. Unexpected, unpredictable AI behavior and current notions of foreseeability

AI behavior is notoriously unpredictable in ways that are significantly out of the distribution of quirks/mistakes seen in human behavior. Take, for example, the self-driving Uber that hit a pedestrian because it did not understand the concept of jaywalking.[58] However, this is not always a bad thing -- AI behavior is often beneficial as a complement to human behavior precisely *because* it is so probabilistically out-of-distribution of expected human behavior. It's ability to think outside the human behavior box makes it excellent at innovation in narrow domains -- this same ability helped AlphaGo defeat the human Go champion,[59] and helped AlphaFold achieve

---

[57] Anthropic. *Project Vend: Can Claude run a small shop? (And why does that matter?)* (https://www.anthropic.com/research/project-vend-1). (research posted by one of the foremost AI companies today, OpenAI's biggest competitor, documenting how their LLM Claude displayed some truly wild and unpredictable behavior when put in charge of stocking an office vending machine)

[58] NPR. Feds Say Self-Driving Uber SUV Did Not Recognize Jaywalking Pedestrian In Fatal Crash. https://www.npr.org/2019/11/07/777438412/feds-say-self-drivinguber - suv - did - not - recognize - jaywalking - pedestrian. National Public Radio 632 (NPR). 2019

[59] David Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: nature 529.7587 (2016), pp. 484–489. (the technical paper introducing AlphaGo, the AI that famously beat the best human player in Go, a game thought to be considerably harder than chess and out of reach of AI systems)

superhuman protein-folding predictions.[60] However, if we want to step outside these highly specialized use-cases and into the long-term goal of human-like artificial general intelligence (AGI), this very capacity for unpredictable, non-human-like behavior becomes the source of the problem.

Certain types of AI models, although cut from the same cloth as all other large models exhibiting unpredictable behavior, have been designed to sound like reasonable humans – ChatGPT for example was trained to produce a "reasonable continuation"[61] of its text prompt. This creates an interesting conflict. AI behavior inherently includes unexpected, unpredictable behavior, and yet it is trained extensively to *sound* reasonable, and it achieves this quite effectively.[62]

As the most sophisticated AI systems are increasingly billed as *generalist* foundation models,[63] the deception of their ability to *sound* reasonable interferes with our legal system's usual heuristics for reasonably foreseeable behavior and responsible decision-making.

There are those who have proposed that an AI should be its own legal entity with its own responsibility for its actions. This is a particularly useful excuse for those AI creators who wish to distance themselves from responsibility for the behavior of their AI systems (such as, in a more amusing example, an AirCanada chatbot that invented customer service policies).[64] Courts, however, have been unsympathetic to this view, and for good reason. Not only does it create perverse incentives for AI creators by absolving them of responsibility, it also creates AI strawmen who are convenient to blame as self-responsible entities, but impossible to punish. In attempting to correctly allocate responsibility and understand foreseeability, we must return, therefore, to the humans on either side of the AI: those

---

[60] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: Nature 596.7873 (2021), pp. 583–589. (technical paper introducing AlphaFold, the AI that famously solved the tremendously difficult problem of protein-folding that human brains still cannot solve)

[61] Stephen Wolfram. What Is ChatGPT Doing…and Why Does It Work? https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/. Stephen Wolfram Writings. 2023. (article explaining ChatGPT's training to produce reasonable-sounding continuations of its prompt)

[62] Cameron Jones and Ben Bergen. "Does GPT-4 pass the Turing test?" *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024. (demonstrating that humans are often fooled into thinking, by conversing with it, that GPT-4 is a human)

[63] Michael Moor, et al. "Foundation models for generalist medical artificial intelligence." Nature 616.7956 (2023): 259-265.

[64] WIRED. Air Canada Has to Honor a Refund Policy Its Chatbot Made Up. https://www. 674 wired.com/story/air- canada- chatbot- refund- policy/. 2024.

creating it, and those interacting with it (or otherwise being downstream of its decision-making).

Given that we know AI behavior can be significantly out-of-distribution of our expected range of human behavior, should this common understanding be baked into our own human notion of foreseeability of AI behavior? If so, how should empirical understandings of human behavior and AI behavior shape what we collectively believe is foreseeable for AI? We can look to technical AI research, cognitive science, and existing legal frameworks to help us construct some answers.

As AI systems are showing major strides towards artificial general intelligence (AGI), we might think that the human assumptions about responsibility for actions should be equally generalizable. However, this would be an enormous miscalculation. A legal solution that ignores the technical realities of individual AI models, and the monumental *differences* between AI behavior and human behavior, would be ineffective and incomplete. It would be wiser for us to take a technology-centric approach to addressing the questions that should underpin the formation of legal standards for foreseeability of and responsibility for AI actions and decisions.

III. WHAT *IS* FORESEEABLE ABOUT AI BEHAVIOR?

A. Non-determinism

Many non-technical researchers express surprise upon learning that AI behavior is *non-deterministic* – that is, and AI can take in the same input in two different instances and produces very different output, despite nothing programmable changing on the inside of the system. In general, for modern deep-learning-based systems (such as LLMs, self-driving cars, and most other systems), non-determinism is a mathematical outcome of the graphics processing unit (GPU) hardware used the train the AI system, and is inescapable—the AI system cannot simply be built to be deterministic.

However, this is only the first of several potential causes of unpredictable output. In addition to the mathematical reality of non-determinism, there are also other reasons for unpredictable output. This includes programmable settings, such as an LLM's "temperature" setting (an external knob provided by some, but not all, LLM creators to control the level of "creativity" in the LLM's response.[65] Each company's internal method of translating a consumer's choice of temperature into a real

---

[65] Zoltán Ságodi, et al. "A Program Synthesis Dataset for LLM Temperature Analysis." *IEEE Access* (2025). (technical paper explaining and demonstrating the use of variable temperature in LLM prompting)

programming change for the model is not publicly revealed, and is therefore poorly understood.[66] Without access to this information, all third-parties can do is study the LLM's *behavior* to try to estimate its level of unpredictability due to temperature. Some AI evaluation efforts have started to attempt this difficult feat to give the public a better understand of the breadth of LLM behavior[67].

B. AI systems will reflect patterns from their training data

It is a fact well-understood from the earliest days of deep learning that an AI system's strength is also its biggest weakness: it learns not from explicit instructions, but from ample training data.[68] There was an era in AI development, before deep learning became the key underlying technology, where computer scientists attempted to hard-code all necessary knowledge into a symbolic AI system.[69] This experiment was endorsed by most of the foremost names in AI research at the time.[70] It failed spectacularly—it turns out that the real world presents unexpected events and developments such that it is impossible to hard-code an AI with all the knowledge it needs. Instead, we turned to machines that learn for themselves, from data, thus understanding the contours and the in-betweens of previously hard-coded concepts, and therefore proving far more robust and human-like.[71] Unfortunately, this also means that the models learned to imitate not only *sounding* human, but having human biases in addition to unexpected AI-specific ones.[72]

---

[66] *Id.*

[67] Noam Kolt et al. "Responsible reporting for frontier AI development." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 7. 2024. (technical paper discussing the emerging field of public-safety oriented technical AI evaluations)

[68] Ángel Alexander Cabrera, et al. "What did my AI learn? How data scientists make sense of model behavior." *ACM Transactions on Computer-Human Interaction* 30.1 (2023): 1-27. (explaining that modern AI models are not programmed in the traditional sense, but rather learn on their own from ample amounts of training data)

[69] Paul Smolensky. "Connectionist AI, symbolic AI, and the brain." *Artificial Intelligence Review* 1.2 (1987): 95-109. (recounting some of the older theories and methods of trying to program an AI from scratch, which were popular in the 60s and 70s but ultimately failed)

[70] Marvin L. Minsky, and Seymour A. Papert. "Perceptrons: expanded edition." (1988). (two of the foremost AI scholars of the time, who wrote a book about how deep learning would not work – they were wrong, but their views stifled the field for a long time)

[71] Jared Kaplan et al. *Scaling laws for neural language models*. arXiv preprint arXiv:2001.08361 (2020). (a seminal recent paper demonstrating that LLM performance scales with how much data they are fed – in other words, the more data, the more powerful the model)

[72] Ninareh Mehrabi et al. "A survey on bias and fairness in machine learning." *ACM computing surveys (CSUR)* 54.6 (2021): 1-35. (survey of technical work documenting racial, gender, and other biases that AI models learn from their human-provided data)

This is a facet of modern models that technical researchers have never been able to rid AI systems of entirely. The previously-discussed example of a computer vision system that learned to classify "wolves" as images with snow in the background[73] is an example of such a statistical pattern being mistaken for the true human-like meaning of a concept. Unfortunately, when these statistical patterns have to do with protected characteristics like race and gender, the results are catastrophic, like the sentencing AI used in the criminal justice system that notoriously gave disproportionately longer sentences to black defendants,[74] ostensibly because it learned from the past decisions of human judges that this was the typical thing to do.

At the very least, AI behavior with respect to protected characteristics must be investigated and externally reported on thoroughly, and any biases corrected to the full extend possible before deployment. However, this is just the tip of the iceberg. AI decisions do not just reflect biases neatly along protected category lines -- they show strange and unexpected biases. Examples observed in large foundation models range from an unusual acuity for counting the number nine[75] and for encouraging self-harm for climate-action purposes.[76] These biases probably derive from the training data in some way, but it is unclear exactly how and why they were learned. More study in this vein is needed, particularly if we continue to give these AI systems significant responsibility in the real world.

C. Humans will not consistently foresee AI behavior

In practice, AI behavior is often strange and surprising in unexpected ways. Characteristics like affinity for particular numbers[77] or a tendency for

---

[73] *See generally,* Roselli et al. (providing several examples of spurious correlations leading to bias in AI)

[74] ProPublica. Machine Bias. https://www.propublica.org/article/machine-bias637 risk-assessments-in-criminal-sentencing. 2016. (investigative reporting exposé of the racially biased AI used in criminal sentencing)

[75] Sunayana Rane et al. "Can Generative Multimodal Models Count to Ten?" In: Proceedings of the Annual Meeting of the Cognitive Science Society. Vol. 46. 2024. (showing that today's most sophisticated AI models cannot reliably count to ten)

[76] Euronews. Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change. https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop600 climate-. (news article explaining the suicide and the AI's alleged role in it)

[77] Sunayana Rane et al. "Can Generative Multimodal Models Count to Ten?" In: Proceedings of the Annual Meeting of the Cognitive Science Society. Vol. 46. 2024. (showing that today's most sophisticated AI models cannot reliably count to ten)

inducing harmful thoughts[78] are things that humans would not ordinarily look for in AI systems that sound so remarkably human, because these are not characteristics of fellow humans who demonstrate a similar level of intelligent language. We are wired to correlate the level of linguistic ability with a general tendency to say true and reasonable things. Models are not so credible, but they do sound credible. Their behavior interferes with our own internal heuristics for credence. In addition, as previously discussed, humans already have known existing biases with respect to foreseeability and responsibility, which are perhaps intuitive and useful when considering human behavior, but are poorly adapted to AI behavior.

Therefore, consumers and other ordinary people should not be responsible for foreseeing AI behavior. The understanding should be, rather, that AI behavior is likely to be both biased and (from a human behavior baseline) often strange. This may often be harmless, but it also may not, and without more information about how the model compares to human behavior and indeed interacts with human behavior, we as a society cannot make informed decisions about where the responsibility for foreseeing AI behavior should truly lie in a given instance.

D. Creators know to expect the unexpected of AI behavior (and consumers do not)

Unlike ordinary consumers, creators of closed-source[79] AI systems have the time, incentive, computational resources, and opportunity to develop intuitions about and even produce internal metrics about their AI's range of behavior, including what kinds of harmful output it can produce. However, creators are seldom incentivized to share this information, or conduct a thorough investigation into foreseeable harms.

When there is clearly a knowledge imbalance, as in this case, the law has traditionally acknowledged this with stricter standards for the party with more knowledge. Take *res ipsa loquitur*, for example, where it is the defendant's responsibility to rebut a presumption of guilt when the

---

[78] Euronews. Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change. https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop600 climate-. (news article explaining the suicide and the AI's alleged role in it)

[79] (closed-source means that the code for the AI system is not publicly available, which is the case for most of the popular models today, including OpenAI's GPT and Anthropic's Claude. The opposite of closed-source is open-source, and there are also a few open-source AI models out there today)

defendant had control over the instrumentality causing the injury.[80] A strict liability products liability regime similarly reflects the understanding that a manufacturer has far more knowledge regarding defects and problems than a consumer could ever have, and meets this unfair information asymmetry with a balancing power of liability. With AI development, and particularly with the excessively broad sweep of trade secret protections around AI systems,[81] we are in similar territory. The creators of closed-source AI systems are by necessity the ones who are most familiar with the behavioral tendencies of their systems, and have the technical sophistication to know that their AI will often produce unexpected, surprising, and perhaps harmful results.

E. For each AI system, creators have the power and responsibility to inform the public of what is foreseeable

Every AI system is different. Modern foundation models have billions of parameters, each trained from data. These parameters, weights and biases, are often randomly initialized and then further optimized using advanced forms of stochastic gradient descent[8283]. The result is that each model, no matter how sophisticated, is different, much like every human brain and mind is different. Unlike human minds, however, AI systems are not grounded in even the most basic shared human concept-level understanding of the world.[84] Therefore, we cannot make the same assumptions and inferences about model behavior as we would make about another human. The only way to figure out what the model has learned is to empirically investigate its behavior.

---

[80] Charles E. Carpenter "The Doctrine of Res Ipsa Loquitur." *The University of Chicago Law Review* 1.4 (1934): 519-535. (a discussion of the *res ipsa* doctrine in torts, for those unfamiliar)

[81] Anna Popowicz-Pazdej. "The proportionality between trade secret and privacy protection: How to strike the right balance when designing generative AI tools." *Journal of Data Protection & Privacy* 6.2 (2023): 153-166. (discussing the current state of sweeping trade secret protections for generative AI systems)

[82] Diederik P Kingma. "Adam: A method for stochastic optimization". In: arXiv preprint 614 arXiv:1412.6980 (2014). (technical paper describing one particular optimizer used in stochastic gradient descent)

[83] John Schulman et al. "Proximal policy optimization algorithms". In: arXiv preprint 646 arXiv:1707.06347 (2017). (technical paper describing a family of optimization algorithms used in practice to train AI)

[84] Sunayana Rane et al. "Concept alignment". In: arXiv preprint arXiv:2401.08672 (2024). (explaining the concept-level misalignment between humans and machines, such that we don't understand the world in the same way)

There must, therefore, be an affirmative duty for those creating closed-source AI systems to thoroughly empirically investigate their behavior, document full efforts to correct harmful tendencies, and illustrate in a way that is absolutely lucid to external third parties, the full breadth of their AI system's behavior. While they choose to make these models closed-source, the models are already affecting public discourse, thought, and the most intimate details of everyday life. Even microwave popcorn comes with ample warning labels; foodstuffs are accompanied by ingredient lists and nutrition information. AI systems are complex modern creations that nonetheless interact with ordinary people on a regular basis, and simply leaving their empirically tested behavior entirely shrouded in secrecy cannot be acceptable if we want to protect human safety. We will never be able to estimate what kind of harm is foreseeable if we do not first know the nature of the AI system, its tendencies and biases, its limitations and blind spots.

IV. WHAT IS AN AI'S CAUSAL UNDERSTANDING OF THE WORLD?

Whether today's most advance AI models can "reason," causally or otherwise, is a subject of intense debate within the computer science and cognitive science research communities. While much has been said both for and against LLM reasoning ability, a realistic observer would conclude that LLMs and foundation models exhibit reasoning-like behaviors under *certain circumstances*, and that those behaviors are brittle and unreliable *under many circumstances*. For almost every reasoning skill studied, the research community has been able to find prompts that lead to successful responses, but also prompts and inputs that lead to wildly incorrect responses. Studies show that LLMs can substitute for human participants in some studies and experiments with allegedly little detriment to the outcome,[85] but others show that they have significantly inferior analogical

[85] Mathew Hardy et al. "Large language models meet cognitive science: LLMs as tools, models, and participants". In: Proceedings of the annual meeting of the cognitive science society. 606 Vol. 45. 45. 2023. (studies using LLMs as replacement for human participants)

reasoning and concept-level understanding skills than those very same humans.[86] [87] [88]

So although there is no single answer to the question "What does an AI's causal understanding of the world look like?", we do at least know that one consistently correct answer is "Not what you would expect of a human's causal understanding of the world." While there may even be enough similarity in output to produce credible, relatively harmless AI survey participants, this understanding does not withstand further rigorous probing at the most fundamental concept-level that we share as humans. For applications where surface-level sophistry is sufficient, where credibly sophisticated-sounding linguistic responses suffice, modern LLMs do nicely. But where such behavior could be harmful if it does not also reflect a thorough underlying conceptual understanding (as it almost necessarily does with humans, who do not have the benefit of learning their rhetorical skills by reading the entire internet, and in whom excellent linguistic ability is usually correlated with above-average intelligence), this mismatch can lead to catastrophic misalignment of human expectations and trust in an untrustworthy model.

A. Modern AI gives the semblance of human-like causal "reasoning"

ChatGPT and other related large foundation models can pass the MCAT, the bar exam, and other tests normally associated with above-average human reasoning skills.[89] [90] And yet, the consensus is that they fail in what many would consider markers of child-level reasoning skills. How can this be? These models are trained to produce responses that seem reasonable to a human counterpart.Trained with RLHF (reinforcement learning from human feedback) in which human annotators rate the AI's

---

[86] Melanie Mitchell. "Abstraction and analogy-making in artificial intelligence". In: Annals of the New York Academy of Sciences 1505.1 (2021), pp. 79–101. (technical paper showing the inability of AI systems to do analogical reasoning)

[87] Melanie Mitchell, Alessandro B Palmarini, and Arseny Moskvichev. "Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks". In: arXiv preprint arXiv:2311.09247 625 (2023). (showing that the GPT models do not perform at human-level on reasoning tasks

[88] Arseny Moskvichev, Victor Vikram Odouard, and Melanie Mitchell. "The conceptarc bench627 mark: Evaluating understanding and generalization in the arc domain". In: arXiv preprint 628 arXiv:2305.07141 (2023) (introducing challenge tests for AI to behave in a human-like way)

[89] Daniel Martin Katz et al. "GPT-4 passes the bar exam". In: Philosophical Transactions of the 612 Royal Society A 382.2270 (2024), p. 20230254.

[90] Vikas L Bommineni et al. "Performance of ChatGPT on the MCAT: the road to personalized and equitable premedical learning". In: MedRxiv (2023), pp. 2023–03.

responses and it then[91] improves iteratively from those ratings,[92] models have been shown to develop traits that make these human annotators instinctively think the AI is being reasonable and impressive, while it may still be inaccurate, misleading, or deceptive under the hood. Model outputs that are obviously untruthful get filtered out by human annotators on whose preferences the RLHF reward model is trained, but that also means that (aside from the imperfect reflection of human preferences inherent in the reward model training process) those untruths which sound generally credible to RLHF annotators often slip through and are positively reinforced.

A model trained to sound reasonable (without any safeguards for whether it is *actually* truthful) is a recipe for many of under-the-hood problems. Yet this is precisely what the RLHF process often creates. We have all experienced the artful hallucinations and misrepresentations that undermine our confidence in an otherwise perfectly reasonable-sounding language model.[93]

But can we really put the burden on ordinary consumers to understand the difference? The untruths that are produced as output in today's most sophisticated AI systems have already passed the credibility test implicit in being reviewed by many RLHF annotators. They are credible-sounding enough to fool lawyers and charm judges.[94,95] It is not a human failing that we are so easily taken in by these untruths; it is rather a very understandable consequence of AI's training process, and our heuristics developed over the

---

[91] (there is additional technical complexity to exactly how this is done, including the training of a second model, known as a "reward model," to mimic human preferences.) *See* Frick, Evan, et al. "How to evaluate reward models for RLHF." *arXiv preprint arXiv:2410.14872* (2024).

[92] Paul F Christiano et al. "Deep reinforcement learning from human preferences". In: Advances 593 in neural information processing systems 30 (2017). (technical paper introducing ideas from modern RLHF)

[93] Yujie Sun, et al. "AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content." *Humanities and Social Sciences Communications* 11.1 (2024): 1-14. (one of many works documenting AI hallucinations and falsehoods)

[94] NYTimes. ChatGPT Lawyers Are Ordered to Consider Seeking Forgiveness. https://www.nytimes.com/2023/06/22/nyregion/lawyers-chatgpt-schwartz-loduca.html. 2023 (lawyers who got in trouble for using ChatGPT in a brief, because it hallucinated citations to cases that didn't exist)

[95] The Guardian. Court of appeal judge praises 'jolly useful' ChatGPT after asking it for legal summary. https://www.theguardian.com/technology/2023/sep/15/court-of-appeal-judge-praises-jolly-useful-chatgpt-after-asking-it-for-legal-summary. 2023. (UK judge using ChatGPT for legal summaries, praising its performance)

course of many interactions with fellow *humans*. These sophisticated models are trained to give the semblance of human-like causal "reasoning," without being able to reliably demonstrate any of the underlying common sense and concept-level understanding that usually come with that ability in fellow humans. It is no wonder that we are so easily fooled, and it should be considered entirely foreseeable that we continue to be fooled while the present training paradigms are in place.

B. Chain-of-thought reasoning and models explaining themselves

With the advent of LLMs that, on the surface, converse quite competently with us in natural language, a reasonable question to ask is whether they can simply explain the reasons for their behavior ("behavior" in this case being their past linguistic interactions). After all, we would afford this explanatory right to a fellow human when challenging their decision-making.

A promising recent line of work investigates this question. It shows that human prompters have the ability to encourage a model to demonstrate better reasoning skills simple by prompting step-by-step, and by asking for explanations along the way in a complicated answer, much as one might do when teaching a child. Called "chain-of-thought reasoning,"[96] this method seemed to reveal hidden capacities for models to reason accurately, including about their own abilities.

However, recent work has shown that these models once again *sound* reasonable in the outputs they produce, but the explanations themselves are unfaithful -- in other words, models don't "say what they mean."[97] To be sure, the method is interesting in that in does seem to elicit more coherently-reasoned responses. However, the responses do not reflect reality, and what is being described, while logically plausible, is often inaccurate. The problems discovered in chain-of-thought reasoning reflect the more fundamental underlying problems of models trained to sound credible. Part of this skill is learning to reason in a coherent-sounding way, whether the reasoning is accurate or not. For now, at least, we cannot rely

---

[96] Jason Wei et al. "Chain-of-thought prompting elicits reasoning in large language models". In: 669 Advances in neural information processing systems 35 (2022), pp. 24824–24837. (technical paper introducing chain-of-thought reasoning and showing that it improves the LLM's reasoning performance)

[97] Miles Turpin et al. "Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting". In: Advances in Neural Information Processing Systems 36 (2024) (recent work showing that models cannot reliably explain themselves)

on models' explanations of their own behavior to understand the causal chains behind their actions.

C. Interpretability research in modern AI systems: the building blocks of responsibility

Where then, do we look for answers? We must go both deeper and broader, to examine not what the models have to say for themselves, but how they behave in practice and even what actually happens under their hoods. This is a notoriously difficult task.

AI systems built using deep learning, which constitute all of the most advanced foundation models today, are notorious for being black boxes -- instead of being hard-coded with discrete knowledge, these systems learn on their own using training data.[98] It is therefore difficult for even the creators of the model to understand *why* a model behaves in a certain way, and what its behavior tells us about its underlying method of modeling the world.

Interpretability research aims to close this gap by looking under the hood, opening up and examine the smallest details of a models' learned weights and biases.[99] It should be noted that the progress in interpretability research is coupled with a healthy dose of reality in the understanding that it has a long way to go before it fully explains model behavior (just as, perhaps, neuroscience has a long way to go before it fully explains human behavior).[100] Nevertheless, interpretability methods give us the best tools we have for probing the insides of the black boxes.

Since the earliest days of deep learning, there have been efforts to illuminate the black box, to varying degrees of success. The early deep learning computer vision systems built using convolutional neural networks (CNNs) were probed using a technique called a *saliency map*, designed to determine which pixels of an input image most caused the output (say, a classification of "bird").[101] Initial analyses showed that the causal pixels indeed corresponded to a human-like concept-level understanding of the

---

[98]Davide Castelvecchi. "Can we open the black box of AI?." *Nature News* 538.7623 (2016): 20. (one of many works describing the black box problem of deep learning based AI)

[99] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: arXiv preprint arXiv:1702.08608 (2017). (providing an overview of the technical field of machine interpretability, and explaining why substantial progress has yet to be made before we understand the internals of AI systems)

[100]*Id.*)

[101] Julius Adebayo et al. "Sanity checks for saliency maps." *Advances in neural information processing systems* 31 (2018). (showing that saliency maps were showing only part of the picture)

bird as an object in the image, providing some human-interpretable views into the black box. While saliency maps and other feature-attribution techniques gained some traction, critiques quickly revealed that even untrained models demonstrated similar tendencies to illuminate human-like concepts, throwing a wrench in the argument that interpretability helped reveal the causal relationships in a black-box post-training.[102]

Similarly, in reinforcement learning (a key component of the methodologies used to train ChatGPT and other state-of-the-art LLMs), there has long existed the credit-assignment problem, the difficult of attributing future outcomes to past experiences.[103] When working with learning machines, their strength of absorbing much from large amounts of training data also lead to commensurate difficulties in properly attributing credit to the right data for a particular output. While we lack this basic understanding of the causality of model behavior, any progress made will be at the scale of individual models or families of models. For example, even if interpretability research cannot predict whether all multimodal models can effectively count as well as a five-year-old, a probing study using interpretability techniques can tell us just how lacking a particular model is in this ability, and how effectively its skills with number concepts can be improved.

Inherent in these efforts, and throughout the many developments in interpretability research, is the question of *why* a model made a decision or produced an output -- in other words, what happened in the inside of the black box to produce an output that seemed reasonable to humans (and indeed, often surpassed human ability on some specialized tasks). As the models became more sophisticated, we began to ask more sophisticated questions about causality -- did the model's underlying causal understanding of the world match our own? In what ways did it differ? Did the models actually have an "understanding" of the world at all? The last question has become increasingly important when considering LLMs that operate only in the linguistic modality -- what does it mean for a system to "understand" the world if it is entirely divorced from any sensory experience? Even if we could understand the causal chains producing its outputs, if we could understand its own internal causal model, would it be meaningful in any way that maps easily to the human experience? Even from a purely practical

---

[102] Julius Adebayo et al. "Sanity checks for saliency maps". In: Advances in neural information 583 processing systems 31 (2018).

[103] Dilip Arumugam, Peter Henderson, and Pierre-Luc Bacon. "An information-theoretic perspective on credit assignment in reinforcement learning". In: arXiv preprint arXiv:2103.06224 (2021). (recent technical work explaining the credit assignment problem)

perspective, this is a difficult question to answer, and one for which the AI research community does not yet have a good answer.

While the field of interpretability research still grapples with more questions than answers,[104] it has developed useful tools and methods for technical analysis that are practically applicable even today, even before we have more general answers about the causality underlying AI behavior at large. While we cannot afford to make broad, sweeping claims about the causal chains underlying AI behavior, and while we cannot make equally broad, sweeping claims about how to accurately foresee AI behavior, we can start with the technical tools that interpretability and cognitive science research give us: the ability to conduct in-depth behavioral and representation-level analyses of model behavior, for each individual model and each anticipated use case. Nutrition labels are different for different brands of the same type of product, enhancing our ability to foresee the consequences of consuming each version of that product; so too should it be for legal transparency requirements for the increasing variety of AI systems.

V. PROPOSED PRINCIPLES FOR ADAPTING THE LEGAL CONCEPT OF FORESEEABILITY TO PROMOTE SAFER AI BEHAVIOR

Having better understood the technical realities of model knowledge and behavior, and the cognitive realities underpinning our own misalignment with model behavior, this section proposes methods to adapt the legal concept of foreseeability to AI behavior.

The first cause of the foreseeability paradox is the dearth of public information regarding testing, evaluation, and behavioral tendencies of the vast majority of AI systems. This creates and environment where ordinary consumers cannot even begin to protect themselves from what they don't know about harmful behavior potential of the AI they are interacting with. The law often addresses harms arising from information asymmetries by creating legal frameworks where responsibilities are commensurate with the underlying information asymmetries. For example, the legal frameworks of *res ipsa loquitor* (to determine negligence in tort doctrine), and of strict liability in product liability doctrine both place the burden of proof with the party that holds a substantial information asymmetry. The presumption is a valuable tool in placing the onus on the party that has the upper hand.

---

[104] Cynthia Rudin et al. "Interpretable machine learning: Fundamental principles and 10 grand challenges." *Statistic Surveys* 16 (2022): 1-85. (explaining the many, many unsolved technical challenges in the field of machine interpretability)

A. Proposed regulatory presumptions

1. Humans will mistakenly attribute human-like qualities to AI systems

The foreseeability paradox has emerged on the back of a fundamental misconception: that ordinary consumers of AI behavior should be able to foresee its harmful behavior, and intervene or act accordingly to prevent it.[105] Instead, the presumption underlying regulatory changes should be that humans interacting with AI systems are likely to attribute human-like qualities to their abilities, including a human-like causal understanding of the world, which is inaccurate to modern AI systems. AI creators should be responsible for incorporating this human tendency into their model safety evaluations, which they must publicly share for transparency regarding a model's human-compatibility.

This proposed regulatory presumption shifts the focus from AI capabilities to the anticipated human response to AI capabilities. As with policy constraints on addictive substances and food labeling, it acknowledges that human behavior has demonstrated a certain proclivity that could lead to harm, and introduces a policy-based safety net for that behavior. Finally, it shifts the legal burden of considering human behavioral responses to the only common-sensical place it *should* be -- with the AI creators who have much the upper hand in the information asymmetry, and who therefore have the *ability,* if not the inclination, to test for, and adapt to, dangerous human misconceptions/responses arising from their AI's behavior.

2. AI creators need regulatory incentives to proactively test for safety and share safety-related information

For closed-source AI systems (those without publicly available code and data), creators of the AI systems are generally the only ones who have sufficient access to the internals to, if they choose, run thorough safety-oriented evaluations. This is the current default, and is likely to remain so unless regulatory incentives emerge to force, and/or reward, sharing of this information.

---

[105] *See, e.g., Benavides v. Tesla, Inc.*, No. 21-cv-21940, 2025 WL 1768469 (S.D. Fla. Jun. 26, 2025) (where defense argued that human in the driver's seat should have stepped in to prevent Tesla Autopilot accident)

Regulatory incentives can take many forms and still address the foreseeability paradox. They can range from statutory and rule-based information-sharing mandates in line with the EU AI Act or California's SB53[106]. They can involve rewards or even, in extraordinary or low-risk cases, safe-harbors for AI creators who go out of their way to be proactive in their information-sharing regarding dangerous AI behavior and safety concerns. They can even remain fairly flexible in leaving it up to the courts to decide exactly what constitutes a best-practice behavioral evaluation of AI, because this is something whose technical definition is likely to change as AI continues to change. All of these regulatory incentives are aligned in their answer to the foreseeability paradox, and their goal of improving the foreseeability of AI harm.

B. Proposed legal presumptions

1. AI behavior is fundamentally different, and more unpredictable, than human behavior

By default, the legal presumption should be that AI behavior is substantially out-of-distribution of expected and foreseeable human behavior – those creating and using AI models should be responsible for giving us data, analyses, and reasons to support or rebut this presumption in their specific cases. This proposed legal presumption aligns incentives with public disclosure and transparency, by ensuring that proactively conducting safety evaluations and making data publicly available is an affirmative responsibility for AI creators. Those who take these proactive steps fulfill their legal burden and are more protected from litigation risk; those who do not are fully exposed. The presumption rewards those who are proactive about determining, and mitigating, foreseeable harms.

2. AI behavior should not be used in areas where decisions have significant possibility of harm to humans

By default, the legal presumption should be that AI behavior should not be used in areas where decisions have significant possibility of harm to humans. Once again, those creating and using AI models should be responsible for giving us reasons to rebut this presumption in their specific cases.

---

[106] *See generally* California SB53 (2025-2026), establishing the Transparency in Frontier Artificial Intelligence Act (TFAIA) (new California AI law regulating large AI models)

This proposed legal presumption, which takes inspiration from the EU AI Act's classification of AI systems' dangers not by their technical sophistication, but by their use case, provides a fail-safe switch specifically for applications where AI misbehavior or unexpected (unforeseeable) human responses to AI behavior, could lead to unacceptable harm. For such use cases, similar to use cases that fall under the "unacceptable risk" category of the EU AI Act,[107] the default is that AI use is inherently not allowed.

These presumptions can be rebutted, as most legal presumptions can. They do not stifle potentially transformative, beneficial AI uses. However, they do align incentives and create affirmative responsibilities that counteract the information asymmetries which are currently a stumbling block to ensuring safe, transparent AI behavior in the wild.

C. Reasons for rebutting presumptions

Legal and regulatory presumptions are usually rebuttable with sufficient evidence. So too in the case of the foreseeability paradox, opportunities to rebut a presumption should form the basis for incentive realignment between the general public and the AI creators who release dangerous products to them.

What forms should these reasons take, when taking into account the evolving nature of technical research and interpretability/human-compatibility data? This section provides some concrete options that effectively help realign incentives by centering regulatory and judicial logic around what the parties do to increase the foreseeability of potential harms – something they currently have little incentive to do.

1. Public release of behavioral analyses laying out the full scope of model behavior

AI behavior is unpredictable by default, but can be probed, characterized, and revealed (akin to peeling back layers of the multi-dimensional onion that is an AI system). Currently, due to the foreseeability paradox, Instead, AI creators should be required to – and rewarded for – conducting and publicly sharing the results of behavioral analyses laying out a detailed picture of model behavior. Results should contain substantial breadth and depth, including anticipated failure modes (or effective deception modes) when interacting with average consumers.

---

[107] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. (also known as the EU AI Act, this is the first comprehensive AI legislation of its kind)

Such a practice would be a significant departure form the current norms. Creators of the main closed-source AI systems (such as Open AI's ChatGPT, Anthropic's Claude, and Google's Gemini) do not publicly reveal any testing or behavioral analyses carried out to understand how the AI's behavior could go wrong. In fact, it is unclear if most creators run such detailed-behavioral analyses at all. With this information publicly available for scrutiny and public response, both halves of the foreseeability paradox would be substantially mitigated: the information asymmetry regarding AI safety/behavior information would be lessened, and the misaligned incentives of AI creators could be recalibrated to reward players who do the responsible thing regarding foreseeability. In what has been described as a competitive "AI arms race,"[108] AI creators may feel like they have no choice but conform to market practices of secrecy. Regulation that litigation that centers the foreseeability paradox can help reverse that detrimental trend.

2. Evidence that the model was tested by independent third parties on a representative sample of the full range of actual use cases, without sugar-coating or obfuscation.

Superior to an AI company's self-released evaluations of their AI systems are the evaluations of third-parties. AI creators have been accused of lying or obfuscating regarding the human-compatibility and safety of their creations.[109] An independent evaluation sets up the type of adversarial opportunity provided by public comment periods in administrative law settings: ideally, it provides the opportunity for a disinterested third-party to provide this crucial safety information. There have recently been concerns regarding private actors influencing or even directly setting their own standards for safety of other products[110]. Information-forcing regulations or rulings can, and should, avoid such a state of affairs for AI safety by proactively requiring both internal and independent third-party technical evaluations of AI systems. These regulations can ensure that the AI was extensively tested on a full range of possible inputs, and without any

---

[108] Paul Scharre. "Killer apps: The real dangers of an AI arms race." *Foreign Aff.* 98 (2019): 135. (engaging with concerns regarding framing AI as an international arms race)
[109] Michael Douglas. "Responsibility of OpenAI for defamation and serious invasions of privacy by ChatGPT." *Communications Law Bulletin* 42.2 (2023): 8-12. (discussing the allegedly secretive and obfuscating behavior of one major AI company regarding the safety of its products)
[110] Joanna R. Schacter "Delegating Safety: Boeing and the Problem of Self-Regulation." *Cornell JL & Pub. Pol'y* 30 (2020): 637. (discussing the problem of Boeing setting its own FAA standards, allegedly resulting in recent safety disasters)

obfuscation of inconvenient or dangerous outputs that could cause public relations damage to the AI creator.

3. Transparent, independently-verifiable data and code for these empirical evaluations

At the heart of the foreseeability paradox is an information asymmetry. Results of a behavioral and safety-oriented technical study are less helpful if we don't know exactly how those results were produced, and which technical judgment calls were made along the way. Transparent, independently-verifiable data and code for any empirical evaluations is critical to remedying the first half of the paradox (this does not necessarily have to be publicly available, if trade secrets or national security are a concern, but they should at the very least be provided on a secure and limited basis to trustworthy independent third parties).[111]

4. Required review process for alternative uses of a vetted model

AI behavior is difficult enough to accurately predict and understand for the use cases the AI system was designed for. The foreseeability paradox is compounded when, as is common in practice, an AI system trained for one task is used for another.[112] In Angwin et al.'s analysis, a team of computational journalists (one of whom has a degree in math) meticulously reverse-engineered the algorithm's decisions over a project spanning several months. They used sparse public data to illustrate racially biased outcomes caused by an AI tool trained to assist in bail decisions, but later marketed and sold for use in sentencing decisions and recidivism risk predictions. This kind of reverse-engineered third-party analysis is an uphill battle and quite impossible for most AI systems, many of which don't have any public data released from which to begin the analysis. Computational journalists are still scarce, and even fewer have enough of a technical foundation to run a research study showing alarming harmful behavior by an AI tool.

The technical AI research community is well aware that best-practices discourage such translational use of AI, but nevertheless AI creators, who aren't subject to public scrutiny or scientific peer review in the same way,

---

[111] Stephen Casper et al. "Black-box access is insufficient for rigorous ai audits." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2024. (explaining several ongoing efforts towards strong AI "audits" by third parties)

[112] Julia Angwin et al. "Machine bias." *Ethics of data and analytics*. Auerbach Publications, 2022. 254-264.

still repurpose AI systems for tasks they should not be used for.[113] Therefore, to remedy this cause of the foreseeability paradox, a review process should be required whenever the model is proposed to be used for a use case different from the one it was originally tested on/for. This review process should serve as an information-forcing protocol that opens the AI's use for public debate, by providing the data necessary to meaningfully discuss the foreseeability of harmful behavior on the new task.

D. A framework of analysis for courts and regulators: Questions to ask when considering whether an AI system's behavior was reasonably foreseeable

Once a dispute related to AI behavior or human-AI interaction makes it through to the litigation stage, what questions must be asked before determining where the foreseeability, and underlying responsibility, should lie? This section draws together the underlying legal analysis into an actionable questioning framework for courts.

The most fundamentally necessary questions are:

- Were the low-level model embeddings tested for common pitfalls, such as race/gender bias, concept-level entanglement, misunderstandings of concepts, and other technical pitfalls known to AI researchers?
- Was a representative sample of the real-world use-cases used to perform these empirical tests beforehand? This sample should be externally verified for representativity.
- Were any unpleasant examples suppressed or removed from the empirical evidence that would have spoken to model (mis)behavior?
- Who provided the test data, and is it available to third parties?

When training models, good data hygiene dictates the need for a "test" data set – a completely "held-out" dataset that the model never sees during training, that is untouched. If the model's performance doesn't match the test set, it is known to have "overfit" and its quality is questioned. The test dataset is generally kept in a *cleanroom* entirely separate from the development teams, and preferably with a third party -- in common machine learning competitions on the platform Kaggle, for example, test sets are never disclosed to any participants building and submitting models, but are

---

[113] *See, e.g.,* the need for Regulation (EU) 2024/1689 at Art. 43 (Article 43 of the EU AI Act, requiring conformity assessments for previous-approved AI systems that change significantly during use or are used for a different purpose than originally intended)

instead used by competition judges to differentiate the truly superior models from those that simply learned to overly mimic the training set.

Completely antithetical to this sound technical principle is the modern practice of AI creators to simple hide away any unpleasant examples of model behavior – this is not good for the company in the long run, because those problems are not properly addressed, and it is (of course) also harmful to the people interacting with the model.

Mandatory disclosure of certain information, such as on warning labels, is not a new idea in product liability doctrine. However, the various disjointed state product liability doctrines have not evolved to the technical realities of modern AI, nor does it solve the foreseeability paradox. For an AI system with billions of parameters (think: neurons) whose encoded knowledge even technical researchers don't understand, a simplistic warning label is vastly insufficient.

The additional complicating factor with "test sets" used on foundation models is that many foundation models have been trained on nearly the entirety of the searchable internet, including most publicly-available testing and evaluation data. This leads to obvious data contamination problems with traditional scientific methods, and forces us to look to empirical behavioral studies of human cognition to glean better methods to evaluate intelligence (and skillful sophistry masquerading as intelligence) when we don't have the luxury of a knowledge *cleanroom*. In other words, one cannot always control what human experimental subjects have or have not seen/read before entering your experiment, and yet cognitive scientists have found excellent ways of probing and characterizing the breadth of human behavior -- AI evaluators can take lessons from this body of work.

1. *Is the use case one where unpredictable behavior could lead to serious harms?*

If the use case one where unpredictable behavior could lead to serious harms, the underlying presumption should be that given AI behavior's unpredictable qualities, the model should not have replaced human decision-making at all. This type of outcome is certainly possible to achieve without explicitly centering new doctrine around foreseeability – such as, for example, Articles 5, 6 and 9 of the EU AI Act , banning AI use entirely in use cases where there is unacceptable risk.[114] However, absent a US federal law or regulatory framework that adopts an explicitly risk-based approach to AI regulation, foreseeability is a viable and potentially farther-

[114] *Id.* at Art. 5-6, 9

reaching alternative that achieves a similar purpose with more room for growth as the technology changes.

2. Can a human override an unreasonable model behavior or decision?

Is there human oversight and accountability in 100% of model use cases? Model creators should be responsible for understanding human theory-of-mind responses to model behavior, and taking at least some basic measures to mitigate them. For example, do the creators acknowledge and build for the fact that humans are likely to believe an LLM when it backs up its claims with (often fabricated) references and (often invented) numbers?[115]

If so, what steps have the model creators taken to make sure the natural, understandable human theory-of-mind responses do not lead to worse outcomes resulting from their models' failure modes? Have they given their users training, with examples, for how their model makes mistakes/invents references, so that the humans can calibrate their understanding of model causality?

These principles provide a framework that courts, engineers, and laypeople alike can use to determine whether AI behavior was foreseeable in a particular instance, and who is responsible for it.

VI. REMEDYING THE FORESEEABILITY PARADOX IN REAL-WORLD AI USE CASES

With this framework, regulators and courts have a foundation from which to begin resolving the foreseeability paradox of AI harms. There are two ways in which foreseeability consistently comes up in a legal analysis: first, humans are expected to foresee the probable consequences of their own actions; second, they are expected, to a certain degree, to foresee the probable consequences of others' actions as long as those actions are equally reasonable and foreseeable.

If, as the Romans thought, foreseeability is a proxy for responsibility, then these are the two ways in which we have, through our accumulated law, deemed that humans are responsible for harms: the harms that they (1) cause and (2) should have foreseen, knowing what they do about the causal consequences of their own actions, the consequences of others' actions, and the normal operation of the world at large.

---

[115] See, e.g. *Garcia* v. *Character Technologies, Inc*., 6:24-cv-01903 (M.D. Fla.), and *Lacey v. State Farm Gen. Ins. Co*., 2025 WL 1363069 (C.D. Cal. May 5, 2025). (court case involving teenage suicide allegedly caused by Character.AI's LLM, and case involving lawyers using ChatGPT's hallucinated citations in a brief submitted to court

In previous sections, this work has illustrated why point (2) is not a facet of the foreseeability analysis that can be easily carried over to a world in which AI behavior, designed to imperfectly mimic human behavior and fly under surface-level human standards, is slowly and silently taking over domains that were previously relegated only to human behavior. This section ties those scientific findings about the true nature of foreseeability, both of human behavior and AI behavior, back to concrete doctrinal principles from which foreseeability emerged.

If we want to solve the foreseeability paradox, what exactly needs to change in the current legal concepts of foreseeability as they apply to AI behavior? If the question asked by Easterbrook's Law of the Horse is implicitly, "Isn't this just the old problem in new packaging? Isn't this something the existing legal canon is well-equipped to deal with?", then what precisely is our answer?

The scaffolding to help answer this question becomes clearer by returning to the simplest restatement of how foreseeability currently plays in legal analyses: first, humans are expected to foresee the probable consequences of their own actions; second, they are expected, to a certain degree, to foresee the probable consequences of others' (individuals', society's) actions. Extending this logic to AI behavior that is now part of the world, it makes sense to ask: first, are humans expected to foresee the probable consequences of AI actions, and second, are AI systems expected to foresee the actions of humans, other AI systems, and the society/world that humans have created around them?

The answer provided by current law is incredibly murky and inconsistent – many of the earliest of these issues are still slowly making their way through the courts, and comprehensive statutes remain rare in the US even at the state level – but our collective understanding of human foreseeability and new emerging data on the inherent unpredictability of AI behavior can help us better understand this.

1. Case A: AI self-driving car

Consider Case A, in which a self-driving car runs over and kills a pedestrian. Such cases have taken the forefront in civil litigation against AI companies, resulting in a recent verdict against Tesla[116] and prior several settlements, including against Uber.[117] In each of these cases, a key question

---

[116] *Benavides v. Tesla, Inc.*, No. 21-cv-21940, 2025 WL 1768469 (S.D. Fla. Jun. 26, 2025)
[117] *Uber settles with family of woman killed by self-driving car*. The Guardian. Mar. 28, 2018. https://www.theguardian.com/technology/2018/mar/29/uber-settles-with-family-of-woman-killed-by-self-driving-car. (describing the Uber self-driving car accident that killed a pedestrian, and the subsequent settlement provided to avoid a lawsuit)

that judge should consider is whether the harm was foreseeable, which relates directly to whether the company in question did everything possible to prevent such harm.

In self-driving cars, massive amounts of proprietary data about the AI system behind the wheel, including specific information about the miscalculation that caused the crash (if indeed it was preventable), are locked beneath the protections that companies put on their software and hardware. Modern AI systems are not explicitly programmed to take certain actions; rather, they learn on their own from large amounts of data, and what exactly they have learned is consistently unclear without time and resources invested in detailed scrutiny. Think of the trained AI system behind the self-driving wheel as an intricate puzzle, which plaintiffs' lawyers cannot easily unpack without vastly more information: What kind of data did the AI system learn from? Were there quality problems with this data? What kind of behavioral nudges were providing during the training process, and did they prioritize safety of the passengers, pedestrians, or both? Perhaps the biggest question of all is, how much did the company know about the tendency for such harmful behavior, and what did they do to prevent it? Given what they knew, should they have put the dangerous product on the market (or the streets) at all?

This may seem like a relatively long series of asks, but each of these technical questions is crucial to understanding whether the AI creator behaved responsibly with respect to the potential for their AI causing harm. AI creators have the choice to devote resources to answering these questions, but often choose not to (perhaps because they lack any incentive to do so). In the case of the Uber self-driving car that ran over a pedestrian, it slowly became clear that the car did not understand the concept of jaywalking: the pedestrian did not cross on a designated crosswalk, so the car ran them over.[118] On the surface, this might seem like a programming error – the programmers simply forgot to include a provision for "jaywalking pedestrians," and *respondeat superior* should apply per usual – but the reality is far more complex than that. In modern AI systems, programmers do not explicitly provide *any* human-understandable concepts to the system (in other words, they no longer do what computer scientists call "hard-coding" or manually entering knowledge into the AI system). Instead, programmers train the AI system with massive amounts of data, and it picks up patterns from that data to learn on its own. Researchers have been working on AI for decades, but it is this ability to learn from massive amounts of data that has given modern AI its impressive, if erratic, high

performance.[119] But this also means that *what exactly the AI system has learned* is unclear to everyone, including the programmers.

This means that even now, the precise cause of the Uber crash is unclear: it could have been due to the AI system not understanding that jaywalkers exist at all, or it could have been that there were no examples of red-headed jaywalkers in its training data, and therefore it didn't learn to extrapolate from the other jaywalking data but instead ran over a red-headed jaywalker. Now we get to the crux of the issue, the key to why an evolved legal concept of foreseeability should be such an important component of the analysis: Should the AI company simply be able to claim that it should not be expected to know about its car's tendency for dangerous behavior towards jaywalkers, or red-headed jaywalkers, or another (unprotected) specific category of people? More broadly, should jaywalking have been a foreseeable external action that *any* self-driving product should be able to safely deal with?

In order to answer these normative questions effectively, courts and regulators alike need more information from the AI company. This is where the necessity and value of the proposed legal and regulatory presumptions constructed in previous sections becomes clear. Starting with (1) "By default, the policy presumption should be that humans interacting with AI systems are likely to attribute human-like qualities to their abilities, including a human-like causal understanding of the world, which is inaccurate to modern AI systems": This regulatory presumption makes it clear that the pedestrian does not have the responsibility of foreseeing that an approaching car may be driven by AI, or that an AI-driven car would not understand jaywalking as a human-driver would. This also means that if they want to shift any responsibility to the human pedestrian, the AI company has to affirmatively produce information that they gave the human enough warning that the car was self-driving (perhaps with a visual label or other flag) *and* that the human should have anticipated that the AI-driven car was going to behave erratically.

This presumption creates an incentive for proactive information-sharing that will be beneficial to human safety. For example, where initially the AI company might be reticent to draw attention to the fact that their car was self-driving and not human-driven, they might now change course in order to show that, in the interest of safety, they made their product's harmful behavior more foreseeable to the humans involved. Similarly, while AI companies previously had no incentive to share testing data that shows that

---

[119] Jared Kaplan, et al. "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361 (2020). (describing the crucial role of big data in making today's AI systems transformative and powerful compared to those of the past)

their self-driving car does not always safely navigate around pedestrians, they may now decide that proactively sharing this information makes them less responsible for future accidents, because this makes the potential harm more foreseeable by proactively sharing data. Lastly, this presumption creates an incentive for the self-driving car company to proactively run tests and adapt their product to real-world human behavior. Whereas previously the AI company could pronounce in vague terms that they conducted "sufficient" safety testing, this presumption forces them to instead be specific in their claims about safety testing specifically for human behavior. It clarifies where the responsibility lies, and therefore reduces uncertainty for the AI company and the ordinary people the product affects.

The second presumption, (2) "By default, the legal presumption should be that AI behavior is substantially out-of-distribution of expected and foreseeable human behavior – those creating and using AI models should be responsible for giving us data, analyses, and reasons to support or rebut this presumption in their specific cases," is oriented towards clarifying the second assumption about human-AI foreseeability. Without this presumption, AI companies, whose products are designed to mimic human behavior, can perpetuate the false belief that the AI system *actually* thinks, or behaves, in a human-like way. A human would not run over a jaywalking pedestrian simply because they don't understand that humans sometimes cross the road without a crosswalk – but an AI system would, and often can't help it. AI designers know this. AI companies know this. Internal tests of the self-driving car would reveal this behavior long before the cars are put on the road. However, the public at large does not understand this, nor do courts or legislators have any data to understand when, how, and how often this happens.

Setting up a legal presumption to ensure that courts think of AI behavior as predictably unpredictable – as foreseeably unforeseeable, if you will, creates an incentive for AI companies to share what their data tells them about the dangers of their own products so that they can beg leniency when things *do* go wrong with the AI behavior.

Under this second presumption, approximately mimicking human behavior *some* of the time is not reason enough to absolve AI companies of responsibility when their products behave in un-human-like ways. Instead, AI companies will be incentivized to clarify when and how their AI systems are likely to deviate from human expectations, and thus to mitigate the risks emerging from these deviations. For example, a self-driving car company would be incentivized to share that their car has not been extensively tested on real-world pedestrian data. A government safety standards bureau can use this data to only permit the AI company to run its cars in certain pedestrian-free zones. Alternatively, using this data, new legislative

standards can develop at the city or state levels where self-driving cars have been legalized, requiring third-party testing or other mandates forcing a detailed study of the AI's behavior before the cars are permitted on pedestrian-dense roads. Some cities or states may even use this data to require further product improvement before the cars are allowed on the road.

One key value of this second presumption is to shift the information asymmetry enough to inform a real public discourse and debate around AI safety standards for all kinds of AI systems: from self-driving cars to automated healthcare AIs. From this point, even if data is falsified, standards are murky, or other gamesmanship is carried out,[120] the expectation to *produce* information that would otherwise never be brought to light will help new standards and caselaw emerge. Case A provides a real-world illustration of the anticipated effects of these presumptions, and indeed the anticipated effects if they are *not* put into place.

2. Case B: AI hiring tool

Consider now Case B, in which an AI tool is used to select or reject applicants in the hiring process. This is another common AI use case that is already well underway in the real world. Alarmingly, AI tools used in the hiring process have been shown to be biased against women applicants, racial minorities, and other groups. The harm caused by this type of AI system is a very different type than that caused by the AI system behind a self-driving car. Usually, the harms include lost employment opportunities, reinforced systemic discrimination, and the intangible value of the loss and disillusionment faced by a qualified applicant denied career progression.

What is not commonly known, however, is that the underlying reason for the unfortunate behavior of many hiring AI systems is the same underlying reason for the erratic behavior of the self-driving car AI: the system has not been explicitly programmed, but rather has learned patterns and incomplete knowledge from large amounts of data. The resulting behavior is designed to mimic human decision-making, but is not human-like under the hood at all, and result in some appalling, puzzling mistakes. For example, the gender and racial bias lawsuits focus only on AI systems that have learned impermissible discriminatory patterns from data. But what about erratic behaviors around non-protected categories or attributes, things we have not sought to protect in the past because until now, there was no

---

[120] Such gamesmanship is to be expected; for example, in *Benavides*, defendants long held that they didn't have the data that would show the cause of the crash, until plaintiffs finally had to hire a forensic expert to recover it. Plaintiffs won a $329 million jury verdict.

reason to assume that widespread discrimination along those categories was taking place?

Take, for example, a hiring AI that has learned never to hire applicants from a particular zip code, never to hire student athletes, and never to hire red-headed interviewees. Such a combination of biases is both believable and at least partially preventable. However, none of these attributes are currently protected under law, so an AI company has no obligation to test for such biases before they put a hiring AI on the market, and plaintiffs have no basis on which to sue for being discriminated against for their zip code or their student athlete status. The erratic, arbitrary set of decision criteria inherent in such a hiring AI are not even known, let alone contestable, under current law.

Once again applying the proposed presumptions, let us start with the presumption that by default, regulators should assume humans interacting with AI systems are likely to attribute human-like qualities to their abilities, including a human-like causal understanding of the world, which is inaccurate to modern AI systems. There is a starting point for remedying this asymmetry. Humans involved in the hiring process – hiring managers, recruiters, team members – would be expected to ascribe human-like abilities to the hiring AI, all while it has hidden, erratic biases against student athletes, redheads, and applicants from a certain zip code. When this presumption is put into place, the responsibility shifts to (1) the companies creating and selling the hiring AI tools, and (2) the companies purchasing and using the hiring AI tools to ensure that the humans involved in the process do not ascribe human-like abilities to the hiring AI.

This could involve providing warnings about potential biases against student athletes, people from Miami, or redheads, so that humans involved in the process may decide to circumvent or override the AI's decision about an applicant from any of those categories. It could also involve detailed trainings on when/how to use the hiring AI, including guidance to only use it as an initial filter, or only use it as a decision-aid. Some companies may choose to always have a human also reviewing the same applicant information fed to the hiring AI, and to use both as an input for decisions. Other companies may choose not to use the hiring AI at all, once they understand the potential risks.

The second presumption, that that AI behavior is substantially out-of-distribution of expected and foreseeable human behavior, and that those creating and using AI models should be responsible for giving us data, analyses, and reasons to support or rebut this presumption in their specific cases, also provides a useful scaffolding for AI hiring tools. Once again, this presumption sets up incentives for the AI companies creating hiring AI tools – the only entity that can provide real answers--to be forthright with

the public about the tool's behavior patterns. To courts, individuals, and customers, the AI company now has a mandate, and a reason, to show that it carried out testing of the hiring AI's behavior on a variety of backgrounds, attributes, zip codes, past activities, and other applicant history. If the AI company finds an unfortunate bias or pattern, it is also in its interest to do everything possible to improve the hiring AI to fix this erratic tendency *before* providing the required safety/testing data to the public. In other words, the presumption forces the AI company to consider, mitigate, and be more open about the foreseeable harms of their hiring AI tool, whereas previously they had only had an incentive to place the hiring AI system on the market as fast as possible. Without such a presumption, hiring AI tools can continue to discriminate in strange and unpredictable ways, causing unseen damage to the hiring process behind the scenes until it is too late.

In both Case A and Case B, the unforeseeable and unpredictable nature of AI behavior has the potential to cause real-world harm (indeed it already has, as evidenced by the corresponding real-world incidents discussed). The cases are vastly different in the type of harm, the use case of the AI, and the potential remedies that courts can apply. As AI use becomes increasingly widespread, each case will lend itself to different specific mitigation strategies depending on the context of the use and the nature of the harm. Nevertheless, in both cases, as different as they are, resolving the foreseeability paradox remains central to a fair and just process: information sharing requirements and incentives to improve the foreseeability and predictability of AI behavior can provide the much-needed space necessary for courts, legislators, computer science experts and the public alike to make real progress in improving AI safety.

CONCLUSION

Easterbrook's "Cyberspace and the Law of the Horse" begins with an entreaty to the lawyers, to let the computer science experts operate in their area of expertise and to avoid creating or proposing new law without extensive computer science knowledge: "I regret to report that no one at this Symposium is going to win a Nobel Prize any time soon for advances in computer science. […] Put together two fields about which you know little and get the worst of both worlds."[121]

AI is a field that affects all of us, though most of us know little about its technical realities. However, if we do not resolve the foreseeability paradox, *even our computer science experts* will be left in the dark about how to

---

[121] Frank H. Easterbrook, *Cyberspace and the Law of the Horse*. University of Chicago Legal Forum 207 (1996).

prevent and mitigate AI harms. AI creators are currently incentivized to release dangerous human-mimicking products into the real world, hiding information about AI's unpredictable, unforeseeable behavior and limiting technical development that could improve the reliability and foreseeability of AI behavior. The computer scientists working to prevent AI harms cannot do their jobs without this information, nor can they inform future AI regulations and litigation with accurate technical knowledge. Resolving the foreseeability paradox can help bridge the technical and legal worlds of AI safety and governance, and go a long way towards flexibly, reliably reducing future AI harms. When it comes to AI, the law has the ability *evolve*, to get the information computer scientists need to inform the public discourse and to make case-by-case decisions about the safety and applicability of AI systems to various aspects of our lives – if we are willing to adapt our legal principles of foreseeability to do it.

As today's state-of-the-art models are built with increasingly closed-sourced codebases, it is the AI system's creators alone who have both the power and the responsibility to approximate what is foreseeable in their individual AI system's behavior. Empirical studies of human behavior make it clear that humans cannot, and should not, be expected to foresee the unpredictable and necessarily unintelligible nature of AI behavior. The legal and policy burden of creating standardized behavioral analyses, of clearly disclosing limitations and dangers, must lie with the only party that has the information needed to do so. This is a legal paradigm seen in frameworks like *res ipsa loquitor*, in strict product liability regimes under tort common law, and in many other instances throughout the regulation of other industries. This work has illustrated how the legal incentives and disincentives should adapt to ensure that those who benefit from AI use must also empower their consumers, their regulators, and the rest of the world with an in-depth knowledge of their AI system's behaviors, including any dissimilarities from expected human behavior. If we reduce this information asymmetry, we can all better foresee the potential harms arising from AI behavior and human responses to that behavior. If we reduce the corresponding incentive misalignment, we can effectively regulate poor foreseeability practices of AI creators no matter how the immediate technology changes in the near future. Creators can make harmful behavior more foreseeable, and therefore more preventable, by investing in studying the interpretability, explainability, and human-compatibility of the AI system, and by making this information publicly available for future AI regulations and litigation. The law can change the incentive misalignment currently preventing them from doing so.

When unexpected AI behavior causes harm, our existing legal frameworks for foreseeability will not help us correctly analyze the issue

and assign blame. This is because AI systems are currently incapable of reliably producing consistent, trustworthy, foreseeable behavior, and AI creators have little incentive to provide information that makes AI behavior more foreseeable, creating a loophole and opportunity for unfettered AI harm. These dual scientific realities, backed by ample data, make it abundantly clear that there is a growing AI safety problem, and that regulators, courts, and even technical researchers need more information about each AI system to determine responsibility for harmful behavior.

The good news is that we can solve the foreseeability paradox. The legal reasoning frameworks centered on foreseeability provide the very scaffolding we need to remodel and reinforce our paradigms to account for the scientific realities of AI behavior and human response. This article provides an path forward to adapt the legal construct of foreseeability to the realities of the AI age. Doing so will help courts accurately determine the responsibility for AI harms, will help producers and consumers alike better understand their respective responsibility in preventing AI harms, and will help policymakers and regulators align incentives such that safer, more human-friendly AI actions are within our reach.

Resolving the foreseeability paradox is not the full solution to effective AI governance, but it is a crucially overlooked and necessary core principle. A transformative technology pushes the law to adapt to keep pace and to ensure justice – resolving the foreseeability paradox is a crucial step towards making this happen.