

Whodunit? Nobody Knows: A Critical Analysis of AI Liability Schemes

Stephen Pearson*

INTRODUCTION

2022–23: Artificial intelligence (“AI”) powered by transformer-based neural networks composed of hundreds of billions of individual parameters operating inside several layers of attention mechanisms and multilayer perceptrons trained for enhanced text prediction through supervised backpropagation becomes widely available for free to an unsuspecting public.

Nobody in 1776 had that on the bingo card.

Or anyone, ever, for that matter.

Neither did Steven Schwartz—a lawyer who employed OpenAI’s newly released generative AI platform ChatGPT (which stands for the rather arcane label “Generative Pretrained Transformer”) to assist in conducting his legal research and later submitted a brief containing fabricated caselaw fresh off the AI engine’s hallucinatory press.¹ Mr. Schwartz, the consternated counsel in question, was hauled into court for a hearing to determine sanctions for his unfounded citations.² His defense? He “just never could imagine that ChatGPT would fabricate cases” and instead assumed that ChatGPT was “a search engine that was using sources [he did not] have access to.”³

* J.D. Candidate, 2026, Sandra Day O’Connor College of Law at Arizona State University. B.S. in Applied Mathematics, Hillsdale College. A huge thank-you is in order to my Note and Comment Editor, Mattheus Thielemann, as well as the 2024–2025 executive board and current staff writers and production editors, especially the excellent Executive Managing Editor, of the *Arizona State Law Journal* for helping to push this project past the finish line. I also want to thank my fiancée, Ana Haynos, and my wonderful family for their support.

1. See Kyle Schnitzer & Priscilla DeGregory, ‘Humiliated’ NY Lawyer Who Used ChatGPT for ‘Bogus’ Court Doc Profusely Apologizes, N.Y. POST (June 9, 2023), <https://nypost.com/2023/06/08/humiliated-lawyer-apologizes-over-chatgpt-court-flub> [<https://perma.cc/MBY5-VF27>]. When an AI system “hallucinates,” it generates an incorrect or completely fabricated answer due to an error in the system’s pattern recognition processes. See *What Are AI Hallucinations?*, IBM, <https://www.ibm.com/think/topics/ai-hallucinations> [<https://perma.cc/H5U9-U7S6>].

2. Schnitzer & DeGregory, *supra* note 1.

3. *Id.*

Without a doubt, “ignorance and carelessness” characterize the conduct in this story.⁴ But before we roll back our recliner and categorize it as another mere example of malpractice, let’s imagine a different scenario. Say a dear old senior citizen—your grandmother or grandfather, perhaps—is running a blog chronicling news updates for a niche topic followed by tens of avid fans. Let us further suppose that he or she is using AI, perhaps unknowingly, to help gather and synthesize sources and stories on this topic. Let us lastly suppose that the AI hallucinates and gives the venerable blogger something falsely but rivetingly slanderous or scandalous. Not knowing or even anticipating that this would happen, our senior posts it, the post goes viral (thanks to sharing and sophisticated search engine algorithms), and the blog’s author quickly becomes the target of defamation lawsuits.

This could be any of us someday, handling an emerging technology that we know very little about and do not understand. Place yourself in Mr. Schwartz’s shoes. Should you, as the end user, be held liable for the AI’s hallucination? Or perhaps the developer should be held liable? After all, the AI itself generated the problematic content. The unfortunate reality, as Mr. Schwartz discovered, is that AI is not perfect and can “commit” actions with negative legal consequences. How should we allocate the liability for such actions? Specifically, in what circumstances should the end user be held liable for actions that, in reality, were committed by the AI?

This Comment argues that to answer this question, we must understand the nature of the seismic shift AI represents. What, in fact, are we dealing with here? AI is a cutting-edge, rapidly advancing technological development that not even its own creators fully understand⁵—a technology projected by some of those creators to surpass human intelligence within the decade.⁶ Different varieties of AI help make hiring decisions,⁷ subvert

4. *Id.*

5. Joseph MacKinnon, *Google CEO Admits He Doesn’t ‘Fully Understand’ How His AI Works After It Taught Itself a New Language and Invented Fake Data to Advance an Idea*, BLAZE MEDIA (Apr. 17, 2023), <https://www.theblaze.com/news/google-ceo-admits-he-doesnt-know-how-his-own-ai-works> [<https://perma.cc/86JX-N6PY>]; 3Blue1Brown, *How LLMs Might Store Facts | Deep Learning Chapter 7*, YOUTUBE (Aug. 31, 2024), <https://youtu.be/9-Jl0dxWQs8> [<https://perma.cc/L9FV-T4PK>] (explaining that the function of multilayer perceptrons inside a transformer architecture is not fully understood).

6. Shirin Ghaffary, *Anthropic CEO Thinks AI May Outsmart Most Humans by 2026*, BLOOMBERG (Oct. 18, 2024), <https://www.bloomberg.com/news/newsletters/2024-10-18/anthropic-ceo-thinks-ai-may-outsmart-most-humans-as-soon-as-2026>.

7. *Mobley v. Workday, Inc.*, 740 F. Supp. 3d 796, 806 (N.D. Cal. 2024) (holding that the employment agency Workday, Inc. could be liable for discriminatory actions taken by its AI hiring tools).

expectations of their users and creators (sometimes with awkward legal consequences),⁸ and ferry us around major cities without direct human supervision.⁹ Most fundamentally and importantly for this Comment, AI is a sophisticated nonhuman entity that is performing tasks previously thought to be only performable by a human. When this man-made creation commits a tort or a crime, who is liable?

This is a largely unpaved area of the law. A uniform framework for analyzing AI liability eludes us.¹⁰ Such vagueness in our legal system is leading to unfortunate circumstances in which prosecutors are unable to directly target clearly abhorrent behavior.¹¹

All of this is to show that we are in uncharted territory here with conflicting policy goals. On the one hand, we wish to prohibit irresponsible and malicious usage of the new capabilities AI offers. On the other hand, we wish to promote the development of this transformative and powerful technology in a manner that is most conducive to societal benefit. A proper solution will set forth a clear liability scheme for actions taken by AI models, a liability scheme with clear boundaries and standards for everyone.

This Comment argues that due to the arcane, unpredictable, rapidly changing, and incredibly complex nature of artificial intelligence models, a negligence shield for end users of AI will function better and promote stronger policy goals than existing forms of assigning liability, including

8. 60 Minutes, *The AI Revolution: Google's Developers on the Future of Artificial Intelligence*, YOUTUBE (Apr. 16, 2023), <https://youtu.be/880TBXMuzmk> [<https://perma.cc/J5ED-M4NA>] (discussing emergent properties); Rose Eveleth, 'My Robot Bought Illegal Drugs,' BBC (July 21, 2025), <https://www.bbc.com/future/article/20150721-my-robot-bought-illegal-drugs> [<https://perma.cc/B85Z-4NC9>] (discussing computer algorithm seized by Swiss police for buying ecstasy pills off the darknet).

9. WAYMO, <https://waymo.com> [<https://perma.cc/5WVT-F474>].

10. See generally Matthew U. Scherer, *Of Wild Beasts and Digital Analogues: The Legal Status of Autonomous Systems*, 19 NEV. L.J. 259 (2018) (providing an overview of different possible AI liability schemes).

11. See Olivia Carville, *Deepfake Pornography Victims Learn There Are No Laws to Fight It*, BLOOMBERG L. NEWS (Nov. 28, 2023), https://www.bloomberglaw.com/bloomberglawnews/litigation/X5FBFBP400000?bna_news_filter=litigation#jcite (showing that generating nonconsensual deepfake pornography has not actually been criminalized) (be warned that this article is not for the faint of heart); see also Isaiah Poritz, *San Francisco Files Nation's First Suit over AI Pornography (I)*, BLOOMBERG L. NEWS (Aug. 15, 2024), <https://news.bloomberglaw.com/litigation/san-francisco-files-nations-first-suit-over-ai-generated-porn> (describing San Francisco's unprecedented legal battle with websites generating nonconsensual deepfake pornography in which the City Attorney is pressing unlawful business practice charges). See generally Michael Goodyear, *Dignity and Deepfakes*, 47 ARIZ. ST. L.J. 931 (2025) (discussing the lack of legal remedies for damaging deepfakes and proposing a solution using the right of publicity).

wild animal, agent-principal, and parent-child liability. A negligence shield will promote usage and development by rendering end users less worried about possible legal consequences from accidentally misusing the technology, but it will still hold liable those who intentionally or recklessly use it to commit a tort or crime. It will also incentivize companies to be on the lookout for and filter out negligent usages by unsuspecting users, which for them is a small tradeoff for being able to reap the benefits of this powerful and lucrative new technology.

Part I sets forth a detailed, albeit brief, summary of AI chatbot and image generation technology. It also describes the current state of AI liability and relevant preexisting legal analogues for AI liability. Part II analyzes these different analogues, explains why they are inadequate, and sets forth the proposed negligence shield framework for AI liability. Part III briefly concludes by reemphasizing the need for such a framework and the inadequacy of current models and proposals.

I. TECHNOLOGICAL AND LEGAL BACKGROUND

Before lawmakers regulate AI and liability stemming from the torts and crimes it assists, they must understand what exactly they are regulating. Its complexity and novelty make it clear that current liability analogues fall short. In addition, it is important to understand the current proposed legal analogues for AI liability to grasp why they fall short. The following two Sections explore the technological and legal backgrounds for AI liability.

A. *Technological Background*

This Section gives a brief but detailed description of the technology underlying the recent explosion of AI relevance, accessibility, and capability. It begins with a short history of neural networks and then sets out a technological primer detailing how these AI programs actually work.

1. Brief History of Neural Networks

The majority of modern AI models depend on a technological structure known as the neural network.¹² First conceived in 1944, the idea of neural

12. See Larry Hardesty, *Explained: Neural Networks*, MIT NEWS OFF. (Apr. 14, 2017), <https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414> [https://perma.cc/G6RR-YZ2H].

networks enjoyed spotty favorability among researchers until recently.¹³ Tomaso Poggio, a prominent researcher at the Massachusetts Institute of Technology (“MIT”), compares the varying levels of interest in neural networks to the different strains of flu viruses: “[E]ach one comes back with a period of around 25 years. . . . In science, people fall in love with an idea, get excited about it, hammer it to death, and then get immunized—they get tired of it.”¹⁴ After the first “trainable neural network” was designed in 1957, neural networks were heavily researched until 1969, but were then pushed to the back burner until the 1980s.¹⁵ They once again fell out of favor in the early twenty-first century, only to make a transformational comeback in the last fifteen years thanks to repurposing graphics processing units (“GPUs”) for neural network processing,¹⁶ a task for which they are particularly well suited.¹⁷ Implementing the neural network architecture on GPUs allowed researchers to multiply the number of layers inside these networks, opening the door to today’s modern systems.¹⁸ Interestingly, a recurring theme in the history of neural networks is researchers’ hesitancy with their “black box” nature: from the 1980s up to the present day, researchers have been unable to determine exactly how these structures work.¹⁹ This accords with the fact that neural networks are “[m]odeled loosely” on the most complicated object in the universe: the human brain.²⁰

2. A Technological Primer

Two prevalent forms of end-user-accessible AI that are possible sources of legal trouble and ambiguity are text-based AI chatbots (such as OpenAI’s ChatGPT, Microsoft’s Copilot, xAI’s Grok, and Google’s Gemini) and AI-

13. *Id.*

14. *Id.*

15. *Id.*

16. *See id.* A GPU is a kind of computer processing component that is designed to run graphics-focused programs. *See What Is a Graphics Processing Unit (GPU)?*, IBM, <https://www.ibm.com/think/topics/gpu> [<https://perma.cc/35XP-MXXP>]. Their “parallel processing” capabilities suit them well for the vast numbers of calculations required for machine learning and artificial intelligence use cases. *See id.*

17. Rick Merritt, *Why GPUs Are Great for AI*, NVIDIA: BLOG (Dec. 4, 2023), <https://blogs.nvidia.com/blog/why-gpus-are-great-for-ai> [<https://perma.cc/C8UD-KAG8>]; *see also* Hardesty, *supra* note 12 (noting that thanks to the video game industry, GPUs were designed and just so happened to fit the computational needs of neural networks nicely).

18. Hardesty, *supra* note 12.

19. *See id.*

20. *Id.*; Julie (Stagis) Bartucca, *The Most Complicated Object in the Universe*, UNIV. OF CONN.: UCONN TODAY (Mar. 16, 2018), <https://today.uconn.edu/2018/03/complicated-object-universe> [<https://perma.cc/54JH-GQ7S>].

powered image generation engines (such as OpenAI's DALL-E 3, Midjourney, or Stability AI's Stable Diffusion). This subsection explores in brief detail how these engines work. In addition, this subsection gives a cursory overview of the concept of emerging properties in AI, which are giving some of its creators concerns. Finally, this subsection briefly surveys the public's reaction to and concerns regarding rapidly developing technology generally and AI in particular.

a. Large Language Models ("LLMs")

LLMs power the most salient examples of artificial intelligence: chatbots such as ChatGPT.²¹ At their core, LLMs essentially use what they have "seen" across the internet to predict the next word, word for word, until the sentence and eventually the whole response is complete.²² According to Stephen Wolfram, "ChatGPT always picks its next word based on probabilities."²³ This prediction algorithm relies on neural networks.²⁴

i) Neural Networks

To conceptualize a neural network, think of an array of nodes arranged in layered columns from left to right.²⁵ In this conception, information flows into the network at the left end and through the network toward the right end.²⁶ To take a well-known example, suppose a neural network is designed to recognize handwritten numerical digits.²⁷ The input would be an image of a handwritten numeral; for the sake of simplicity, let this image be digitized as a black-and-white image with each pixel of the image having a brightness value between 0 and 1.²⁸ This array of pixel brightness values is fed into the neural network; the goal is to have this data flow through the different nodes and then produce an output: a correct identification of the

21. Stephen Wolfram, *What Is ChatGPT Doing . . . and Why Does It Work?*, STEPHEN WOLFRAM: WRITINGS (Feb. 14, 2023), <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work> [<https://perma.cc/RV9F-Y6X4>].

22. *Id.*

23. *Id.*

24. *Id.*

25. See Fangfang Lee, *What Is a Neural Network?*, IBM, <https://www.ibm.com/topics/neural-networks> [<https://perma.cc/E45V-C8N8>] (providing an illustration of a simple "deep learning" neural network algorithm).

26. This is known as a "feedforward network." See *id.*

27. See 3Blue1Brown, *But What Is a Neural Network? | Deep Learning Chapter 1*, YOUTUBE (Oct. 5, 2017), <https://youtu.be/aircAruvnKk> [<https://perma.cc/U5JV-7DUJ>].

28. See *id.* for this very helpful example.

handwritten numeral.²⁹ Each node represents a function; that is, each node receives inputs (either the raw data initially inputted into the network or the outputs from the previous layer),³⁰ manipulates the inputs, and sends the output to each node in the next layer.³¹ This function would include various parameters that can be controlled by the network designer to affect the output of the network.³² Once the input flows through all the layers, the network produces an outcome: final probabilities for each possible numeral the original image could have been (zero through nine), of which the numeral with the highest probability is returned to the user as the identification of the inputted image.³³

For a more grounded, detailed example, suppose there exists a “neural network” consisting of a single node—in other words, one function that takes input data to predict an output.³⁴ To envision a single node working in this way, imagine making a simple decision of whether or not to buy a particular car. If the output of the node is greater than zero, the customer buys the car. Imagine that the customer has three different criteria: performance, style, and reliability. For any particular car, each criterion is rated on a scale from one to ten. Imagine that for this particular car—in this case, assume that the car is a 2011 Audi R8 V10 Spyder manual—performance is rated at an eight, style is obviously at a ten, and reliability might be a four. Therefore, the input data flowing into the node would comprise the following array:

$$x_1 \text{ (performance)} = 8$$

$$x_2 \text{ (style)} = 10$$

$$x_3 \text{ (reliability)} = 4$$

Suppose further our hypothetical customer values style the most, values performance to some degree, and does not much value reliability since he is

29. *Id.*

30. If the node is on the leftmost layer of the neural network, it would receive the overall input (the digitized image of the handwritten numeral); if the node is one of the inner layers, it would receive the outputs of the previous layer.

31. See Lee, *supra* note 25. Formally, this function can be denoted as $f(x) = \sum_{i=0}^n w_i x_i + b$, where n is the number of nodes in the previous layer (for the first layer, n would be the number of pixels), w_i is a specific weight associated with the current node, x_i is the output of a node in the previous layer, and b is a bias number specific to the current node. See *id.* The output of this function would then be normalized to a number between zero and one using a function such as the sigmoid function, which the node would then send on to all the nodes in the next layer. See 3Blue1Brown, *supra* note 27.

32. See Lee, *supra* note 25.

33. 3Blue1Brown, *supra* note 27.

34. The following example is based off an extremely helpful illustration in Lee, *supra* note 25.

buying a rather impractical supercar. Then he might weigh each criterion as follows:

$$w_1 \text{ (weight for performance)} = 6$$

$$w_2 \text{ (weight for style)} = 10$$

$$w_3 \text{ (weight for reliability)} = 2$$

Finally, assume that the customer will not buy the car unless the product of all these weights and criteria sum to at least 150. Therefore, the number 150 will be subtracted from the overall sum as a “bias value.” In the end, the result would look something like this:

$$\text{Output value} = x_1w_1 + x_2w_2 + x_3w_3 - \text{bias value} = 8*6 + 10*10 + 4*2 - 150 = 6.$$

In this case, since the output number is greater than zero, the customer will buy the car.

Modern, AI-capable neural networks are built on layers upon layers of these kinds of nodes, all taking input from either the user (the first layer) or the preceding layer (all the internal layers), processing it, and outputting data to the next layer.³⁵

ii) Backpropagation

But how do researchers come to the particular weights and biases assigned to each node? This is where “training” the neural network comes into play.³⁶ Training basically consists of inputting data into a neural network, recording the output, and then “telling” the network what you were actually looking for.³⁷ This takes place via “backpropagation” using a “cost function.”³⁸ At its simplest level, the cost function essentially measures the difference between the expected or desired output of the network—what you told the network you wanted—and the actual output it generated.³⁹ Each input the network is given results in a different cost function output (i.e., the difference between the actual output and the

35. Lee, *supra* note 25.

36. See Diletta Goglia, *Backpropagation for Dummies*, MEDIUM (July 27, 2021), <https://medium.com/analytics-vidhya/backpropagation-for-dummies-e069410fa585> [<https://perma.cc/V6GG-5KHD>].

37. Andreas Stöffelbauer, *How Large Language Models Work: From Zero to ChatGPT*, MEDIUM (Oct. 24, 2023), <https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f> [<https://perma.cc/JR2Y-85NN>].

38. Goglia, *supra* note 36.

39. *See id.*

desired output), so the overall cost function (“OCF”) constitutes an average of all the cost function outputs for different inputs into the network.⁴⁰

The end goal is to minimize the output of the OCF.⁴¹ Using multivariable calculus techniques, one can find the individual weights and biases for each node that will result in the lowest output for the OCF.⁴² Taking the partial derivatives of the OCF with respect to every individual weight and bias number present in every single node gives the gradient of the cost function.⁴³ The gradient gives the information needed to adjust the weights and biases to produce a lower cost function output.⁴⁴ This process is then repeated to minimize the cost function output.⁴⁵

In the example above about buying the Audi, assume that although the model predicted that the customer would buy the car, he did not (meaning that the output *should* have been less than zero). To train the node to be more accurate, the weights would have to be adjusted so that the ultimate output value was less than zero. Perhaps reliability was actually worth more to the customer than performance, in which case the weights should have been more like this:

$$w_1 \text{ (weight for performance)} = 1$$

$$w_2 \text{ (weight for style)} = 10$$

$$w_3 \text{ (weight for reliability)} = 8$$

This would lead to the following output:

$$\text{Output value} = x_1w_1 + x_2w_2 + x_3w_3 - \text{bias value} = 8*1 + 10*10 + 4*8 - 150 = -10.$$

Since the output value is now less than zero, the newly trained model now correctly predicts that the customer will not buy the car. This is an extremely simplified example of backpropagation training.

iii) Generative Pretrained Transformers

Chatbots like ChatGPT are based on a specific kind of neural network called a generative pretrained transformer (“GPT”).⁴⁶ In this special kind of neural network, the inputs are the user-generated prompts entered into the

40. See 3Blue1Brown, *Gradient Descent, How Neural Networks Learn | Deep Learning Chapter 2*, YOUTUBE (Oct. 26, 2017), <https://youtu.be/IHZwWFHWa-w?si=INxhar2TNKnqjKY> [<https://perma.cc/2ADM-JL3V>].

41. Goglia, *supra* note 36.

42. *Id.*

43. *Id.*

44. *Id.*

45. *Id.* This process is known as “gradient descent.” *Id.*

46. Ivan Belcic & Cole Stryker, *What Is GPT (Generative Pretrained Transformer)?*, IBM, <https://www.ibm.com/think/topics/gpt> [<https://perma.cc/DB3U-T9NL>].

chat, and the output is the text generated by the network.⁴⁷ The user's inputted prompt is analyzed by the model and broken up into "tokens," which are each represented mathematically as vast lists of numbers called vectors (in OpenAI's GPT-3, each token vector consisted of 12,288 numbers).⁴⁸ This list of vectors is then fed into the network, which comprises a series of layers called transformers.⁴⁹

Each transformer layer is composed of two "steps" for treating the inputted list of vectors: (1) the attention step and (2) the feed-forward step, the latter of which is otherwise known as a multilayer perceptron.⁵⁰ The attention layer performs an operation that gathers context from all the other relevant vectors and transfers it to each vector, essentially allowing the "words" to talk to one another and induce context-sensitive changes in the vectors as they flow through the GPT.⁵¹ Essentially, the attention step enables the network to "retrieve information from [other] words in a prompt."⁵²

The modified vectors then pass through the multilayer perceptron, which can be thought of as "a database of information the model has learned from its training data."⁵³ It is composed of a vast "hidden layer" that contains—at least in GPT-3—49,152 nodes, and it uses these nodes to adjust the values of the inputted vector.⁵⁴ Researchers theorize that these multilayer perceptrons contain the network's general knowledge about the world.⁵⁵ It uses this information to attempt to "tentative[ly]" predict the next word, and that information is then passed on to the next transformer layer.⁵⁶ Ultimately, however, the function of these multilayer perceptrons is not fully understood by researchers and scholars.⁵⁷ In any event, the vectors originally inputted by the user as a prompt pass through a series of these transformer layers (each including attention and multilayer perceptron

47. *See id.*

48. Timothy B. Lee & Sean Trott, *A Jargon-Free Explanation of How AI Large Language Models Work*, ARS TECHNICA (July 31, 2023), <https://arstechnica.com/science/2023/07/a-jargon-free-explanation-of-how-ai-large-language-models-work/> [<https://perma.cc/4QFF-JHSJ>].

49. *See id.*

50. *See id.*; 3Blue1Brown, *Transformers, the Tech Behind LLMs | Deep Learning Chapter 5*, YOUTUBE (Apr. 1, 2024), <https://youtu.be/wjZofJX0v4M?si=hMUqnKGk0Vz9Elyd> [<https://perma.cc/ZFK9-HC3V>].

51. *See Lee & Trott, supra* note 48.

52. *Id.*

53. *Id.*

54. *See id.*

55. 3Blue1Brown, *supra* note 5 (discussing conclusions of Google DeepMind researchers on the topic).

56. *See Lee & Trott, supra* note 48.

57. *See* 3Blue1Brown, *supra* note 5.

steps) until the network predicts the next word (or, more technically speaking, token) of the output.⁵⁸

GPTs of this sort are trained by inputting large amounts of readable text found all around the internet: through backpropagation, the network essentially attempts to predict the next word in the passage, compares the predicted result to the actual result, adjusts the parameters inside the attention and multilayer perceptron layers, and continues its attempts at prediction.⁵⁹

b. Diffusion Models

Diffusion models are another kind of neural network most famously used in AI-powered image generation software such as OpenAI's DALL-E 3, Stability AI's Stable Diffusion, or Midjourney.⁶⁰ Diffusion models use an LLM to analyze and convert the input text into an array of vectors.⁶¹ Training the diffusion model is similar to the basic neural network architecture described in subsection I.A.2.a: using a cost function to represent "misalign[ment]" with the user-inputted prompt, the neural network's parameters are trained to associate certain kinds of images with certain kinds of prompts and more accurately generate appropriate images.⁶² The model is developed by conducting forward diffusion and reverse diffusion.⁶³ In forward diffusion, developers add Gaussian noise to an image until it is pure static.⁶⁴ Then, in reverse diffusion, the model learns to reconstruct the original image out of that pure static.⁶⁵

The actual image generation process consists of applying that training to a text prompt inputted by the user and encoded by an LLM.⁶⁶ Starting with an image of random noise, the model predicts noise to be subtracted from

58. See Lee & Trott, *supra* note 48.

59. See *id.*; see also *supra* Section I.A.2.a.ii.

60. See Kemal Erdem, *Step by Step Visual Introduction to Diffusion Models*, ERDEM (Nov. 1, 2023), <https://erdem.pl/2023/11/step-by-step-visual-introduction-to-diffusion-models> [<https://perma.cc/72PH-E76P>]; Suhaib Arshad, *How Do Diffusion Models Work? Simple Explanation: No Mathematical Jargon, Promised!*, TOWARDS AI (June 3, 2024), <https://towardsai.net/p/l/how-do-diffusion-models-work-simple-explanation-no-mathematical-jargon-promised> [<https://perma.cc/DDS3-J5RC>].

61. Arshad, *supra* note 60; *supra* Section I.A.2.a.

62. See Arshad, *supra* note 60. For a description of cost functions as measurements of the gap between the desired output and the actual output of the network, see *supra* Section I.A.2.a.ii.

63. Arshad, *supra* note 60; Erdem, *supra* note 60.

64. Arshad, *supra* note 60.

65. See Arshad, *supra* note 60; Erdem, *supra* note 60.

66. See Arshad, *supra* note 60.

the original noise in accordance with the text prompt and its training, and it does this iteratively until the output image is formed.⁶⁷

c. Emergent Properties

A brief word on emergent properties is appropriate. Emergent properties are commonly understood to be properties of complex organisms that cannot be ascribed to any particular component of the organism, but rather arise out of all the components working together.⁶⁸ In 2023, Google executives believed that their AI engine—called Bard at the time—had evinced emergent properties.⁶⁹ As an example, one executive explained that with no training dedicated to learning the language Bengali, Bard learned Bengali completely after a few prompts in the language.⁷⁰

Chalking up these unexpected capabilities to emergent properties is not without its critics.⁷¹ In 2023, a research effort by Rylan Schaeffer of Stanford University showed that so-called emergent properties might appear due to a shortcoming in measuring model capabilities.⁷² Essentially, Schaeffer wondered if these emergent properties were in reality developing in a more linear fashion undetected by standard capability measurement techniques.⁷³ Forbes contributor Andréa Morris describes the then-standard method for measuring AI capabilities: “If the AI wasn’t perfect, an emergent ability wasn’t clocked until it was perfect.”⁷⁴ She explained that this method of measurement “might make a new skill look like it emerged sharply and unpredictably, when the AI was actually improving at tasks at a gradual rate.”⁷⁵ Nevertheless, proponents of the original measurement methods still maintain that emergent properties are a fact of artificial

67. See Erdem, *supra* note 60; *AI Image Generation Explained: Techniques, Applications, and Limitations*, ALTEXSOFT (July 9, 2023), <https://www.altexsoft.com/blog/ai-image-generation> [<https://perma.cc/Q7NJ-FA7X>].

68. David Galas, *Systems Biology*, ENCYC. BRITANNICA (Nov. 10, 2025), <https://www.britannica.com/science/systems-biology>.

69. MacKinnon, *supra* note 5.

70. *Id.*

71. See Katharine Miller, *AI’s Ostensible Emergent Abilities Are a Mirage*, STANFORD HAI (May 8, 2023), <https://hai.stanford.edu/news/ais-ostensible-emergent-abilities-are-mirage> [<https://perma.cc/FWY3-BLRE>].

72. *See id.*

73. *Id.*

74. Andréa Morris, *AI ‘Emergent Abilities’ Are a Mirage, Says AI Researcher*, FORBES (Sept. 29, 2023), <https://www.forbes.com/sites/andreamorris/2023/05/09/ai-emergent-abilities-are-a-mirage-says-ai-researcher/>.

75. *Id.*

intelligence.⁷⁶ The important takeaways from this debate are that AI exhibits some suddenly apparent capabilities, that these capabilities pose concern for those worried about the rapid development of AI, and that researchers disagree on whether or not these capabilities are as consequential as generally thought.⁷⁷

d. Public Concerns

Regardless of whether emerging properties are being correctly measured, the public is concerned about these complex technologies. According to a Pew Research survey, U.S. adults demonstrate a considerable lack of knowledge of the ubiquity of AI-powered services and devices, and they are nervous about the rapid rise of this technology.⁷⁸ This survey also shows that a lower awareness of interactions with AI-powered services and devices correlates with being “more concerned than excited about increased use of AI in daily life,” suggesting that Americans are afraid of what they do not know.⁷⁹ A 2020 survey indicated that 60% of respondents were concerned that technology was developing too rapidly, with trust in technology declining 4% globally year over year.⁸⁰ To make matters worse, engines like ChatGPT are rapidly becoming more “human-like” as they become more sophisticated and complex, obscuring their shadowy, mysterious internals and perhaps lulling their users into a false sense of familiarity and trust, with the end users assuming they are working with something they do understand when in fact they do not in the slightest.⁸¹

76. *Id.*

77. Miller, *supra* note 71.

78. Brian Kennedy et al., *Public Awareness of Artificial Intelligence in Everyday Activities*, PEW RSCH. CTR. (Feb. 15, 2023), https://www.pewresearch.org/science/2023/02/15/public-awareness-of-artificial-intelligence-in-everyday-activities/?gad_source=1&gad_campaignid=22378837192&gbraid=0AAAAA-ddO9GfhH4eyMuxQ4cExvY4VtHIT&gclid=CjwKCAjwpcTNBhA5EiwAdO1S9tftOaPDBaYgPyUnd2wv-8s3LOFlqYk7uHnnBrKwTd5jxFGPL0ZSghoCLJQQA_vD_BwE [https://perma.cc/X2R9-SGH2].

79. *Id.*

80. Rae Hodge, *60% of People Worry That Tech Is Moving Too Fast, Study Finds*, CNET (Feb. 25, 2020), <https://www.cnet.com/tech/tech-industry/global-trust-in-technology-declining-report-says/> [https://perma.cc/HVV8-BDYG].

81. See Jacob Krol, *OpenAI's GPT-4o ChatGPT Assistant Is More Life-Like Than Ever, Complete with Witty Quips*, TECHRADAR (May 13, 2024), <https://www.techradar.com/computing/artificial-intelligence/openai-gpt-4o-chatgpt-assistant-is-more-life-like-than-ever-complete-with-witty-quips> [https://perma.cc/BPZ8-F6HU].

B. Legal Background

This Section examines examples of current and proposed schemes for AI liability. It begins by outlining some current examples of AI liability litigation to emphasize that society is in dire need of a uniform AI liability scheme. This Section then briefly addresses the idea of granting legal personhood to AI and explains why this Comment will not analyze it in depth. Following that, it sets forward the basics of product liability, both wild and domesticated animal liability, liability for children, and agency liability. Lastly, this Section concludes with a brief explanation and example of negligence shields.

1. The Current State: AI Pornography, Federal Legislation, *Mobley*, and *Cohen*

Congress's partisan-seized lawmaking engine haltingly lurches toward a semblance of a solution for one specific instance of liability for AI malfeasance in the DEFIANCE Act.⁸² This bill would create a means of civil legal redress for victims of nonconsensual AI-generated deepfake pornography against the users generating the pornography.⁸³ Furthermore, in May of 2025, President Donald Trump signed the bipartisan TAKE IT DOWN Act into law, criminalizing the dissemination of “nonconsensual intimate visual depictions” and specifically targeting AI-generated deepfakes.⁸⁴

Unfortunately, however, many day-to-day victims of AI-assisted malfeasance find themselves without a legal remedy,⁸⁵ and a uniform framework for analyzing AI liability eludes us.⁸⁶

82. See Kate Tenbarge, *The Defiance Act Passes in the Senate, Potentially Allowing Deepfake Victims to Sue Over Nonconsensual Images*, NBC NEWS (July 24, 2024), <https://www.nbcnews.com/tech/tech-news/defiance-act-passes-senate-allow-deepfake-victims-sue-rcna163464> [<https://perma.cc/H2VL-AU6A>] (creating a means of legal redress for victims of nonconsensual deepfake pornography); S. 3696, 118th Cong. (2024).

83. Tenbarge, *supra* note 82; S. 3696, 118th Cong. (2024).

84. TAKE IT DOWN Act, Pub. L. No. 119-12, 139 Stat. 55 (2025) (codified at 47 U.S.C. § 223(h)); Barbara Ortutay, *President Trump Signs Take It Down Act, Addressing Nonconsensual Deepfakes. What Is It?*, AP NEWS (May 20, 2025), <https://apnews.com/article/take-it-down-deepfake-trump-melania-first-amendment-741a6e525e81e5e3d8843aac20de8615> [<https://perma.cc/8JDR-2VXF>].

85. Carville, *supra* note 11; see also Poritz, *supra* note 11.

86. See generally Scherer, *supra* note 10 (providing an overview of different possible AI liability schemes).

In August of 2024, San Francisco City Attorney David Chiu filed a lawsuit against the owners of sixteen AI-powered websites that generate pornography from user-uploaded photos of women and girls.⁸⁷ This lawsuit is unprecedented, and Chiu is pursuing prosecution under a theory of unlawful business practices.⁸⁸ As of June 2025, ten of the sixteen websites have been forced to stop operating—a huge step in the right direction.⁸⁹ In another case, a high school student in Levittown, New York, used one of these websites to generate disturbing, nonconsensual explicit imagery of his classmates, and the only legally punishable offense was an actual nude photo of a minor taken without consent stored on his phone.⁹⁰ None of the nonconsensual explicit imagery he generated with the assistance of AI was actually illegal because no law in New York actually criminalized it.⁹¹

In situations like the above, where an end user intentionally manipulated software to achieve a heinous result, courts may be open to the idea of preventing liability from running to the developing organization. At least one court has stated in dicta that intentional misuse of otherwise innocuous software does not create liability for the software’s manufacturer.⁹² In *Mobley v. Workday, Inc.*, the court stated that a software vendor could not be held liable for discriminatory actions taken by its software if the software was intentionally manipulated by an end user to achieve a discriminatory result.⁹³

That was not quite the case in *Cohen v. United States*, however.⁹⁴ In *Cohen*, the court held a lawyer who mistakenly submitted cases manufactured by Google Bard to be negligent—“perhaps even grossly negligent”—but nevertheless not liable for sanctions thanks to the lawyer’s “good faith” belief that the cases had come from a reputable source.⁹⁵ This provides a good example of possible judicial willingness to grant some end users leeway with regard to negligence when dealing with AI.

87. Poritz, *supra* note 11.

88. *Id.*

89. Luz Pena, *SF Shuts Down 10 of the World’s Most-Visited Websites Using AI to Generate Explicit Content*, ABC7 NEWS (June 2, 2025), <https://abc7news.com/post/deepfake-porn-san-francisco-shuts-down-10-worlds-most-visited-websites-using-ai-generate-explicit-content/16638231/> [<https://perma.cc/6Q3X-N8RB>].

90. Carville, *supra* note 11.

91. *Id.*

92. *Mobley v. Workday, Inc.*, 740 F. Supp. 3d 796, 807 (N.D. Cal. 2024).

93. *Id.*

94. *United States v. Cohen*, 724 F. Supp. 3d 251 (S.D.N.Y. 2024).

95. *Id.* at 254–55, 258.

2. Legal Personhood for AI

Occasionally, scholars make a functionalist argument for granting AI models legal personhood: if it walks like a human and talks like a human, it must be treated like a human.⁹⁶ For example, Katherine Forrest notes the historical tendency to deny legal personhood to groups just as human as those granted legal personhood; she analyzes AI evolutionarily and functionally and compares this question to the legal status battle for minority groups and women.⁹⁷

However, a short analysis exposes the dangers and uncertainties of such a liability scheme. In order to have legal autonomy, the AI must have legal responsibility.⁹⁸ At the moment, AI models are heavily supervised and controlled by their designing entities.⁹⁹ As Matthew Scherer points out, if society were to grant legal personhood and responsibility to AI models, the entities formerly responsible for them would feel much less of a need to rein them in.¹⁰⁰ The key question is this: is society prepared to let these AI models loose on their own in our real-world economy and communities? If so, then legal personhood can and should be on the table. Otherwise, it is a fool's errand. Moreover, as currently designed, they cannot stand alone: they do not have standing, they do not have bank accounts in their names, and they do not, in many cases, have autonomous bodies, so enforcing judgments and sentences against them would be difficult conceptually as well as practically.

Hence, this Comment does not closely consider the possibility of legal personhood for AI systems. It instead focuses on other possible analogues for AI liability. These are each outlined below.

96. See, e.g., Katherine B. Forrest, *The Ethics and Challenges of Legal Personhood for AI*, 133 YALE L.J. 1175 (Apr. 22, 2024) (making the argument that we should be vigilant and open to the possibility of granting AI legal personhood); Jason Zenor, *Endowed by Their Creator with Certain Unalienable Rights: The Future Rise of Civil Rights for Artificial Intelligence?*, 5 SAVANNAH L. REV. 115 (2018).

97. See Forrest, *supra* note 96, at 1179–98.

98. *Personal Autonomy*, STANFORD ENCYCLOPEDIA OF PHILOSOPHY (Summer 2010 ed.) (“Most of us want to be autonomous because we want to be accountable for what we do, and because it seems that if we are not the ones calling the shots, then we cannot be accountable.”).

99. See *Introducing GPT-5*, OPENAI, <https://openai.com/index/introducing-gpt-5/> [<https://perma.cc/TQS4-A39B>] (advertising “a sharp drop in hallucinations” as well as enhanced abilities to recognize and mitigate subtly “malicious” prompts or unsafe situations).

100. Scherer, *supra* note 10, at 263.

3. Product Liability

Under product liability schemes, a company or any other entity in a supply chain (for example, developers, manufacturers, or retailers) can be held liable for harm caused to a party by a product the entity passed along to the consuming public.¹⁰¹ Strict liability, negligence, and breach of warranty provide three avenues under which the plaintiff can seek a judgment.¹⁰² Under strict liability, the defendant entity is liable for any damage caused by the product to the parties, provided that: (1) when sold, the product was unreasonably dangerous; (2) the consumer would receive the product without that dangerous condition being ameliorated; and (3) the product in fact caused the injury.¹⁰³ In a negligence claim, any person “reasonably affected by the product” may seek a judgment, provided that the ordinary elements of negligence are met: (1) the defendant owed a duty of a minimum standard of care toward the plaintiff; (2) the defendant breached that duty of care; and (3) the breach was the actual and proximate cause of the injury to the plaintiff.¹⁰⁴ Lastly, an injured plaintiff may pursue a breach of warranty claim under express or implied warranty, provided that there is an actual sales contract between the defendant and the plaintiff.¹⁰⁵ The specifics of express and implied warranties are outside the scope of this Comment.

4. Animal Liability

Under animal liability schemes, owners of animals can be held responsible for the actions the animal takes.¹⁰⁶ In this scenario, “[t]he owner of an animal is one who has legal title to the animal” or “gives evidence of ownership.”¹⁰⁷ This liability is generally separated into two classifications: liability for domesticated animals and liability for wild animals.¹⁰⁸ For animals “classified as wild”—i.e., “wild or dangerous domestic animals”—the owners bear strict liability.¹⁰⁹ For domesticated animals, strict liability only applies if the owner knew or had reason to know that the domesticated

101. ROBERT J. GUTE, *PRODUCT LIABILITY CLAIMS, DEFENSES, AND REMEDIES* (2025).

102. *Id.*

103. *Id.*

104. *See id.*

105. *Id.*

106. Scherer, *supra* note 10, at 281–82.

107. 2 *PERSONAL INJURY—ACTIONS, DEFENSES, DAMAGES* § 6.04[1] (2025).

108. Scherer, *supra* note 10, at 282.

109. 2 *PREMISES LIABILITY—LAW AND PRACTICE* § 8C.02 (2024). The injured party cannot have been unlawfully on “the premises on which these animals are kept.” *Id.*

animal had dangerous propensities.¹¹⁰ This can be demonstrated by showing that the owner was aware of the animal committing a dangerous act before—a legal principle known as the “one free bite” rule—since it gives the owner of a domesticated animal one dangerous incident before becoming strictly liable for his animal’s actions.¹¹¹

5. Liability for Children

The liability of parents for the actions of their children is more nuanced.¹¹² At birth, children are not legal persons, instead only possessing passive rights such as freedom from abuse or the right to an education.¹¹³ At this stage, parents bear responsibility for the child’s actions.¹¹⁴ As children grow older, develop physically and mentally, and exercise more discretion and control over their own actions, parents’ rights over their children and liability for their children’s actions decrease.¹¹⁵ Eventually, children achieve full legal personhood, and their parents are released from liability.¹¹⁶

6. Principal-Agent Liability

According to the Restatement (Third) of Agency, a fiduciary principal-agent relationship arises when the parties mutually manifest assent that the agent will act on behalf of the principal and under the principal’s control.¹¹⁷ Restricting the large body of principal-agent law to what is relevant for AI liability, Scherer highlights three specific scenarios of principal-agent liability from the Restatement as particularly pertinent:

- 1) The agent commits tortious conduct while acting with the principal’s “actual” or apparent authority or with the principal’s assent;
- 2) The principal’s negligence in selecting, training, retaining, supervising, or otherwise controlling the agent results in an injury to a third party while the agent was acting for the principal;

110. *Id.*

111. Scherer, *supra* note 10, at 282.

112. *See id.* at 283–84.

113. *Id.* at 283.

114. *Id.* at 284.

115. *Id.*

116. *Id.*

117. § 1.01 (AM. L. INST. 2006).

- 3) The principal attempts to delegate to someone else a duty to protect another, regardless of an actual principal-agent relationship, and the delegatee fails in this duty.¹¹⁸

Scherer posits that implementing a combination of these principles can solve the problem of assigning AI liability.¹¹⁹

7. Negligence Shields

Negligence shields of the general kind proposed below are not an entirely novel idea. A kind of this negligence shield is exemplified in *United States v. Cohen*, in which the court granted a negligence shield of sorts to a lawyer who mistakenly submitted cases manufactured by Google Bard, even if the lawyer was “perhaps even grossly negligent.”¹²⁰ Because the lawyer believed in “good faith” that the cases had come from a reputable source, the court did not impose sanctions.¹²¹ For the purposes of this article, then, a “negligence shield” for AI actions is defined as a prohibition against holding the end user liable for actions committed by AI due to the mere negligence of the end user.

II. ANALYSIS

Clear as mud? Exactly. If you did not understand much if anything about how AI models work and how the proposed legal frameworks would apply to them, that is precisely the point. The core of the argument against the above-proposed legal frameworks is that AI is not like any of their subjects. It is complex, sophisticated, and yet mechanistic. It can also be extremely deceptive for the uninformed, as Mr. Schwartz discovered after he attempted to quiz ChatGPT on its own reliability.¹²² Google’s own CEO, Sundar Pichai, described AI as possessing some “black box” characteristics—meaning that not even the researchers “understand” exactly what occurs within its shadowy neural layers.¹²³ This “opacity is still

118. *Id.* §§ 7.04–06, 7.08; Scherer, *supra* note 10, at 287.

119. Scherer, *supra* note 10, at 287–90.

120. 724 F. Supp. 3d 251, 254–55, 258 (S.D.N.Y. 2024).

121. *Id.* at 255, 258.

122. The AI responded that all of the cases it had cited were real and available on legal research databases. *Mata v. Avianca, Inc.*, 678 F. Supp. 3d 443, 458–59 (S.D.N.Y. 2023).

123. Scott Pelley, *Is Artificial Intelligence Advancing Too Quickly? What AI Leaders at Google Say*, CBS NEWS (Apr. 16, 2023), <https://www.cbsnews.com/news/google-artificial-intelligence-future-60-minutes-transcript-2023-04-16/> [<https://perma.cc/7Q7H-S86V>].

unsettling to theorists,”¹²⁴ raising questions as to whether we should start implementing traditional frameworks on a development not even fully understood by its creators. This Part will counter the above-proposed analogues for AI liability and then propose a new, original solution: a negligence shield for end users.

A. Counters to Relevant Existing Models

As mentioned above, the common theme throughout the below critiques of these existing analogues is that AI simply is not an ordinary product, animal, agent, or child and should not be treated like any of them.

1. Product Liability

Of the alternatives currently available, a product liability scheme may be the best society can do at the moment for assigning liability for injuries caused by AI.¹²⁵ Brandon Jackson, a proponent of a product liability scheme for AI wrongdoing, points out that it would cover many good rationales for assigning liability to the developer.¹²⁶ These include a developer’s failure to train the AI using high-quality data or a failure to warn consumers “of dangerous consequences.”¹²⁷ Furthermore, it could conceivably have the advantage of lowering societal risks from AI: if liability for AI wrongdoing were to be assigned to its developers through product liability, it would incentivize developers to ensure the AI was safe to use before it ever enters commerce.

However, while product liability presents a decent framework for attributing liability to the manufacturer, it presents significant problems that ultimately outweigh any benefits that may result from its implementation. As Jackson himself points out and as discussed above, a product liability claim must establish some sort of defect in the product or breach of duty by the manufacturer: “fault [must be] discernible.”¹²⁸ Unfortunately, it is hard to discern fault inside a technology so complex and opaque as AI.¹²⁹ And if something more akin to a strict liability scheme were applied, perhaps via

124. Hardesty, *supra* note 12.

125. Brandon W. Jackson, *Artificial Intelligence and the Fog of Innovation: A Deep-Dive on Governance and the Liability of Autonomous Systems*, 35 SANTA CLARA HIGH TECH. L.J. 35, 58 (2019).

126. *Id.*

127. *Id.*

128. *Id.*; see discussion *infra* Section I.B.3.

129. Jackson, *supra* note 125, at 58.

res ipsa loquitur,¹³⁰ AI development would be chilled, possibly to the detriment of society.¹³¹ Furthermore, AI system developers generally do not “program [them] from the ground up”—components of the AI software source from several different external origins, ensuring that fault is even harder to find.¹³² Furthermore, AI is not just a standard product such as a wood chipper—it has an ability to make mostly autonomous, inscrutable decisions and act independently based on its own internal “reasoning.”¹³³ Treating it as merely a standalone product would oversimplify the calculus and may not give the flexibility needed, especially as this technology evolves in the future. “[S]trict liability is too blunt an instrument for a technology as inherently malleable as learning A.I. systems,” Scherer states.¹³⁴ For that fundamental reason, product liability is an inadequate liability scheme for AI wrongdoing.

2. Animal Liability

With regard to applying animal liability schemes to AI systems, the “owner” of the AI could be either the developer or the end user. Hilyard Nichols proposes establishing the developer as the owner of the AI system and applying the domesticated animal “first bite” rule to AI injuries: developers are allowed one strike before they are held strictly liable for the harms the AI commits.¹³⁵ However, this approach discounts the existence of end users. Developers do not simply own and control the AI completely as a pet owner owns and controls a dog: if it were not for consumers opening up the AI programs on their personal devices, no harms would be committed by AI. The first bite rule does not take this into account since it assumes

130. The Latin translates to “the thing speaks for itself.” *Res Ipsa Loquitur*, CORNELL L. SCH.: LEGAL INFO. INST., https://www.law.cornell.edu/wex/res_ipsa_loquitur [<https://perma.cc/UM9Z-N8NJ>]. The phrase stands for the legal principle that the accident was of such a kind that usually does not occur without the negligence of the defendant, leading to a presumption of the defendant’s negligence. *See id.*

131. *See* Jackson, *supra* note 125, at 60–61; Alicia Lai, *Artificial Intelligence, LLC: Corporate Personhood as Tort Reform*, 2021 MICH. ST. L. REV. 597, 618–20 (2021) (stating that strict product liability would deprive society of many benefits conferred by AI since it would chill innovation in the field as manufacturers become more cautious); Scherer, *supra* note 10, at 281.

132. Scherer, *supra* note 10, at 280–81.

133. *See* Jackson, *supra* note 125, at 58 (“[A]utonomy is significant, and culpability cannot be easily discerned.”).

134. Scherer, *supra* note 10, at 289.

135. *See* Hilyard Nichols, *The First Byte Rule: A Proposal for Liability of Artificial Intelligences*, 15 WM. & MARY BUS. L. REV. 189, 209–10 (2023).

developers control their AI models as simply and directly as pet owners control their dogs. Furthermore, Nichols himself points out that, unlike a dog bite, it may be hard to quantify and categorize harms committed by AI (and thus it would be difficult to say that the employer has notice of all dangerous issues once the AI has committed an injury, undermining the core assumption of the first bite rule).¹³⁶

On the other hand, assuming that the end users are the legal “owners” of the AI model for the purposes of AI liability under the first bite rule is too reductionist for such a sophisticated technology. While requiring that the end user defendant have previous knowledge of propensity for misbehavior may partly rectify this issue, it still too easily assigns liability to the end user for something that is considerably more conceptually complex and difficult to understand than your indoor cat. For example, senior citizens might not even be able to recognize signs of AI misbehavior such as hallucination (a phenomenon in which AI models “hallucinate” and claim falsities as factual¹³⁷), at least not in the same way that they would recognize patterns of dangerous behavior apparent in a dog exhibiting violent tendencies. Dogs and cats, after all, have been around for thousands of years; widely available LLMs have hardly been around for three.

Treating AI models as wild animals makes these problems even worse. The structural issues of animal liability still apply for assigning ownership to the developers, but now they are held strictly liable for a system they are not even operating as the end user. If the owner is assumed to be the end user, the result is simply untenable. Assuming that all end users of AI make use of the technology knowing that the technology is inherently dangerous and liable to commit torts and crimes is a rather unjustified leap in logic for two reasons.

First, the end-using public likely does not have enough knowledge and information to know this, at least not with any confidence. Using ChatGPT is not the same thing as letting a bear or a Komodo dragon loose in the neighborhood without a leash. ChatGPT is facially much less threatening; whatever risks it does pose to the larger community are far less obvious (and harder for the user to recognize) than those of your average apex predator. Furthermore, these complex technologies are intentionally marketed as being streamlined and user-friendly, obscuring their complex

136. *See id.* at 212–13.

137. *What Are AI Hallucinations?*, *supra* note 1.

nature and effectively discouraging their users from reading too deep into their internals.¹³⁸

Second, it is also bold to assume for legal purposes that mainstream AI models actually pose the same inherent, tangible risks as wild animals. Many AI developers do in fact place quite a premium on building safety precautions into their models,¹³⁹ rendering them of a different nature from a wild animal. An AI model is not going to break out of your computer and attack the neighbor's toddler (at least, not yet). Therefore, it does not make sense to assign liability for AI models as we would for an aggressive venomous pet snake.

3. Liability for Children

First of all, as Scherer points out, adopting the graduated legal responsibility and liability scheme used for parents and children for AI and its developers and end users presents the personhood problem: “the end result of such a legal framework would be legal personhood for A.I. systems.”¹⁴⁰ Furthermore, Scherer makes the valid point that this would result in the developers eventually being freed from all responsibility for the actions of a system they had more control over than parents do over their children.¹⁴¹

That aside, once again, children have been around for a long time, just like pets and animals. Although they may be puzzling and inexplicable at times, they are much better understood by the general populace than the inner workings of an LLM neural network. We have thousands of years of recorded human history to analyze the behavior of individual humans, not to mention generational wisdom handed down from parents to children or between friends. For example, if a mother is facing a difficult situation with her child, she can seek advice from her parents, her friends, or her mentors, all of whom are likely to have had experience raising or interacting with children. OpenAI does not have the same luxury when facing issues with training and managing ChatGPT. There is no generational wisdom

138. See, e.g., *Claude Sonnet 4.6*, ANTHROPIC, <https://www.anthropic.com/claude/sonnet> [<https://perma.cc/95FF-9ND2>] (“Anyone can chat with Claude using Sonnet 4.6 on Claude.ai, available on web, iOS, and Android.” (emphasis added)).

139. See, e.g., *Our Approach to AI Safety*, OPENAI (Apr. 5, 2023), <https://openai.com/index/our-approach-to-ai-safety> [<https://perma.cc/6YL3-AK5M>].

140. Scherer, *supra* note 10, at 285.

141. *Id.* Scherer points out that AI developers can actually design the “physical components” and internal processes of their AI models, something parents cannot do with their children. *Id.*

regarding the training of LLMs. And if end users are considered for the purposes of this analogue to be the “parents” of AI, the same problems present themselves. Therefore, we cannot expect developers or end users of AI to hold the same liability for AI models as parents do for their children. While developers may train AI, neither end users nor developers live with AI (yet), nor are they responsible for raising AI to be a good, contributing member of society (yet). The application of the parent-child liability paradigm to developers or end users of AI models is thus simply not appropriate.

4. Principal-Agent Liability

As set out above, a fiduciary principal-agent relationship arises when the parties mutually manifest assent that the agent will act on behalf of the principal and under the principal’s control.¹⁴² First, it is clear that this relationship does not arise between the AI model and the developer for torts and crimes committed by end users while using AI: the AI is acting on behalf of the end user, not the developer. Therefore, this possible liability scheme must be analyzed assuming the end user is the principal and the AI is the agent. Turning to the three scenarios laid out in subsection I.B.6 demonstrates why principal-agent liability is not a helpful liability scheme for AI wrongdoing.

First, assume the principal is liable when the agent commits tortious conduct while acting with the principal’s “actual” or apparent authority or with the principal’s assent. For end users to avoid liability under this scheme, they would have to anticipate and somehow block any erroneous direction the AI might go. In other words, implementing this scheme would require an expectation that end users are always aware of every danger that AI represents and could effect in order to avoid liability. Not only that, but they would also need to possess the capability to prohibit the AI from taking a harmful action. Is such an expectation realistic? Not even those designing AI are fully aware of its capabilities and dangers, or even fully knowledgeable about how these systems work. It seems unreasonable to expect that ordinary end users would have that knowledge. Furthermore, even if they did have that knowledge, they are not in fact in control over the AI in the same way a principal is in control over an agent. In reality, in addition to the end user, there is often a behemoth organization such as Google that, in reality, possesses much more direct control over the model

142. See discussion *supra* Section I.B.6.

than does the end user. Therefore, this principal-agent scenario does not seem to provide a satisfying answer to the question of AI liability.

Second, assume the principal's negligence in selecting, training, retaining, supervising, or otherwise controlling the agent results in an injury to a third party while the agent was acting for the principal, placing the liability on the principal. To apply this to the end user of AI models would be a fool's errand. This form of liability presumes that the principal in fact knows that an agent exists and is working for him. For our purposes, this would imply that end users actually have knowledge of when exactly they are using AI and what exactly it is. However, fatally for this scenario, many end users have no idea how AI works or even when they are using examples of it.¹⁴³ It is unreasonable to expect these end users to know how to supervise and control an LLM or diffusion model, not to mention the fact that they are largely irresponsible for training the model. They have neither the power nor the knowledge to be reasonably held liable under this scenario.

Third, assume the principal attempts to delegate to someone else a duty to protect another, regardless of an actual principal-agent relationship, and the delegatee fails in this duty. While this scenario may reasonably cover situations in the future where AI is delegated protective responsibilities, at the moment, this scenario does not generally apply to end users since they do not usually delegate to an engine like ChatGPT the responsibility of looking out for someone else. Therefore, this scenario cannot form the core of a proposed AI liability framework.

This brief analysis of proposed liability frameworks for AI liability shows that current analogues fall short in the face of this incredibly complex, unprecedented, and transformative technology.

B. Proposed Solution: A Negligence Shield for End Users of AI

This Section sets forth the high-level argument for a completely new scheme for AI liability, describes the proposed solution's framework, and justifies the added liability assigned to the organizations developing the AI models.

143. See Kennedy et al., *supra* note 78.

1. Why Not an Analogue: Bird's Eye View

Why not implement a legal analogue for AI liability? The answer is simple: we have not seen anything like it before.

As shown above, these models are unimaginably complex. They are “inscrutable” and “unpredictable” in their shortcomings.¹⁴⁴ Scherer points out that they present unique issues regarding autonomy, foreseeability, control, and “programmed” as opposed to “intended” goals.¹⁴⁵ Indeed, Scherer specifically states that “the opacity of AI systems may make courts hesitant to blame the end user of an AI system that causes harm to a third party.”¹⁴⁶ David Vladeck hypothesizes that a “duty to train” may evolve, meaning a developer must competently train, for example, an AI-powered autonomous car before it is released into the wild.¹⁴⁷ Mihailis Diamantis states that “black box algorithms” that “write themselves,” such as neural networks, are “far too complex for any human intelligence to deconstruct or comprehend.”¹⁴⁸ These are all unprecedented developments, and they deserve an unprecedented liability scheme.

This is reflected in the public's trepidation concerning the rise of AI and other rapidly evolving technologies.¹⁴⁹ This nervousness of the public indicates that the public itself believes that society is confronting a new and mysterious technology that does not conform well to any analogue. Courts and lawmakers should not shrug off these concerns by forcing an old liability scheme onto this new phenomenon. People tend to be afraid of what they do not know,¹⁵⁰ and a legal liability framework for AI should take this into account by creating a new system.

The upshot is this: AI is not a human. AI is not an animal. AI has only been around in its present form for a little over three years (since the release of ChatGPT took the world by storm in late 2022).¹⁵¹ The general public does not understand how it works, and even its creators do not fully understand it. This calls for a fresh solution.

144. Margot E. Kaminski, *Regulating the Risks of AI*, 103 B.U. L. REV. 1347, 1372 (2023).

145. Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 363–68 (2016).

146. *Id.* at 373.

147. See David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117, 140 (2014).

148. Mihailis E. Diamantis, *Vicarious Liability for AI*, 99 IND. L.J. 317, 327 (2023).

149. See discussion *supra* Section I.A.2.d.

150. See *id.*

151. See *Introducing ChatGPT*, OPENAI (Nov. 30, 2022), <https://openai.com/index/chatgpt/> [<https://perma.cc/MTD7-6TJP>].

2. Proposal: A Negligence Shield for End Users of AI

This Comment proposes a negligence shield for the end users of AI, which would apply when an end user is sued for the malfeasance of an AI model acting under his or her supervision.

Under this proposal, when an AI model commits a tort or a crime while being operated by an end user, the end user is not responsible for it if he or she is only negligent. In other words, in order to hold an end user liable for the tortious or criminal actions of an AI model, the government or the plaintiff must prove a minimum mental state of recklessness.¹⁵² For torts, then, there must be more than a simple breach of the reasonable standard of care.

Importantly, this should be a rebuttable shield. It should be able to be overcome by proving one or more of the following scenarios:

- 1) The defense is shown to have technical expertise or knowledge in the field of artificial intelligence;
- 2) The defense is shown to have practical experience due to previous experiences of the specific AI model in question committing tortious or criminal actions of a similar kind under the defense's supervision (a modification of the first bite rule);¹⁵³ or
- 3) The defense is shown to have purposefully used an AI model clearly and obviously designed to be dangerous (such as AI pornography generators).

The first scenario would ensure that the negligence shield would not extend beyond the reasoning for it. The shield's primary justification lies in the fact that AI is new, opaque, and difficult to understand. If AI is shown not to be so for the defense, the negligence shield should not apply.

The second scenario is related to the first and also serves an important policy goal. Due to their previous unfortunate experiences with AI malfeasance, end users with such practical experience should not be able to hide behind their ignorance of AI systems in deflecting liability. Furthermore, this exception incentivizes end users with such experience to use more caution when using AI in the future.

152. For the purposes of this article, recklessness is taken to mean that the possibility of the tortious or criminal action actually crossed the mind of the individual and the individual ignored that possibility.

153. This practical experience—which could be actual or constructive knowledge—could be proven by the following fact patterns, for example: (1) the defense was notified in the past that a specific tort or crime was committed by the AI in the course of the defense's use of the AI, (2) the defense actually knew that such a tort or crime was committed in the course of the defense's use of the AI, or (3) the defense ever faced adverse legal action for actions taken by the AI.

The third scenario ensures that the negligence shield will not be weaponized to provide cover for using AI that is intrinsically and intentionally dangerous.

Outside of these three scenarios, this negligence shield would promote an important policy goal. It would promote widespread use of AI, thus driving this important industry forward, by offering end users a legal security blanket. The result will be that they will not be constantly looking over their shoulders, worried that they might be held responsible for an AI model they do not really understand.

3. The Developers Should Be Held Liable Instead

The negligence shield set forth above should deflect the liability onto the organizations developing and taking the AI models to market. If the developing organization is different from the organization taking the model to market, the marketing organization should be held liable.¹⁵⁴ This is because such an organization is responsible for actually placing this technology in the hands of uninformed consumers, making it the lowest-cost avoider in the supply chain, as it may simply train the end users on how to use the technology.

Holding the developing and marketing organizations liable may seem like a tremendous and unprecedented leap, but it is less outrageous than it sounds and serves important policy goals. Furthermore, there is reason to believe that more stringent liability schemes will not actually drastically chill development, as these organizations may be incentivized by competition and the promise of untold profits to continue developing AI models regardless of the liability scheme.¹⁵⁵

a. More Prudent Decision-Making Is Possible

Charlotte Tschider points out that although neural networks are complex and opaque, they are in fact ultimately designed by humans, and specific responsible development choices are possible.¹⁵⁶ Occasionally, these practices and concerns are deliberately sidestepped, ignored, or avoided.¹⁵⁷

154. Disputes over this would likely be addressed in contracts between the developing and marketing organizations, which would likely include indemnification agreements.

155. See Nichols, *supra* note 135, at 214.

156. Charlotte A. Tschider, *Beyond the "Black Box"*, 98 DENV. L. REV. 683, 690–93 (2021).

157. See Sigal Samuel, *OpenAI As We Knew It Is Dead*, VOX (Sept. 26, 2024), <https://www.vox.com/future-perfect/374275/openai-just-sold-you-out> (noting that in recent

OpenAI, before it lost its nonprofit character, even held back from releasing GPT-2 to the public for fear of end user misuse.¹⁵⁸ Therefore, forcing companies to ensure their models have more guardrails to protect against negligent misuse is not too tall of an order.

b. They Have the Control

The developing organizations alone understand, as much as anyone is able to understand in the context of modern AI models, what they are doing and are able to control their models. They are the least-cost avoiders of accidental AI harm. They have the background knowledge and control the technology, so they are in the best position to be able to protect against negligent misuse. Furthermore, they do not even *want* other people to be in their position to fix these problems: they would likely rather keep the specific architectures under wraps to protect their trade secrets.

On the other hand, imposing negligence liability on the end users would come with a greater cost to society. For a typical end user to avoid accidental AI harm, she would have to avoid more conduct than the developing companies would have to prohibit. This is because of the information and expertise gap between individual end users and development organizations. An end user seeking to avoid accidental AI harm would likely try to avoid any instructions to the AI that may conceivably result in the “black box” taking a tortious or illegal action. This means that many uses of AI, especially experimental uses, by individuals would likely be chilled for fear of negligence liability—typical end users cannot possibly hope to have the confidence and expertise needed to make these technical judgments. Conversely, developers are more likely to know more precisely what kinds of prompts and end user actions are likely to result in tortious or illegal conduct, and they can thus tailor their guardrails and prohibitions more narrowly to avoid these pitfalls.

c. An Easy Out

Lastly, although the negligence shield would deflect liability for negligent misuse by end users onto the developing or marketing organization, the organization has several easy ways to solve this problem.

years, OpenAI has championed commerciality over responsibility, recently abandoning its nonprofit status altogether); Carville, *supra* note 11 (noting Stability AI’s refusal to “stress-test” their image generation software for abuse potential before marketing it for profit).

158. Cat Ellis, *AI News-Writing System Deemed Too Dangerous to Release*, TECHRADAR (Feb. 15, 2019), <https://www.techradar.com/news/ai-news-writing-system-deemed-too-dangerous-to-release> [https://perma.cc/JTB8-WJHV].

First, the companies could actually remedy the situation and bulletproof the models against accidental, negligent misuse. While this may sound like an impossible task, such cases of misuse will likely source mostly from end users who are not familiar with the workings of AI. These kinds of snares will likely surface quickly (for example, hallucination, fictitious legal cases, or copyright infringement) and exhibit patterns across the user base that are easily filtered out or at least able to be remedied with obvious disclaimers appended to the AI's response.

Additionally, the negligence shield could easily be rebutted by compelling users to take a mandatory tutorial when signing up to use the AI model. Such a tutorial would give the user enough knowledge of the dangers and shortcomings of the AI models to fulfill the knowledge requirement of the first presumption-rebutting scenario set out in subsection II.B.2. In addition, this would have the added benefit of forcing the company to thoughtfully consider exactly what it is making available to the public and how to teach the public about its shortcomings before anyone uses the model. Additionally, this kind of honesty would encourage the developing organizations to market a more ethical, responsible product if only for a positive impact on public relations. This solution would not chill any development of these models; rather, it would encourage transparency, prudence, and intentionality among the developing organizations.

Finally, the negligence shield does not make the organization liable for manipulation or malicious use of its models. Like firearm manufacturers, AI development companies manufacture possibly dangerous tools and make them available to the public; that does not mean they are liable for bad actors intentionally or recklessly employing them to harm others. This is already somewhat reflected in the DEFIANCE Act, which grants a federal civil remedy to victims of nonconsensual AI pornography generation but does not hold the development organizations liable.¹⁵⁹ The court in *Mobley* has already held that deliberate illegal manipulation of computer software lacking decision-making authority to achieve discriminatory results does not render the software manufacturers liable.¹⁶⁰ This reasoning could easily be applied to AI development companies.

159. See Tenbarge, *supra* note 82 and accompanying discussion; see also Scherer, *supra* note 10, at 287–89 (advocating a scheme where AI developers are not held liable for malicious modifications made to their systems, which are then used to create harm); Daniela Glavaničová & Matteo Pascucci, *Vicarious Liability: A Solution to a Problem of AI Responsibility?*, 24 ETHICS & INFO. TECH. 28, 37 (2022) (stating that a company that sells an AI-powered robot that is then repurposed by a “hacker” and goes on to commit a tort or crime is not liable for the misuse of the robot).

160. *Mobley v. Workday, Inc.*, 740 F. Supp. 3d 796, 807–08 (N.D. Cal. 2024).

III. CONCLUSION

In the end, AI is not a human. AI is not an animal. AI is not really an agent, and it is not a child. Therefore, the liability schemes for traditional analogues will not work in assigning liability for AI's tortious or criminal actions. AI is vastly complicated, and even those who designed it do not fully understand it. Pinning liability for negligence on the consumers who use AI would be tantamount to expecting them to know the ins and outs of such an opaque system, and this would choke the widespread adoption of a crucial modern technology. Pinning liability for negligence on the organizations that design and market AI would incentivize them toward transparency, conscientious development, and education of their products' users, resulting in positive effects for the use and advancement of AI in society. For both policy and logistical reasons, a rebuttable negligence shield for end users of AI models seems to be the best path forward.