BOTS BEHAVING BADLY: A PRODUCTS LIABILITY APPROACH TO CHATBOT-GENERATED DEFAMATION

Nina Brown[*]

INTRODUCTION

Within two months of its launch, ChatGPT became the fastest-growing consumer application in history with more than 100 million monthly active users.[1] Created by OpenAI, a private company backed by Microsoft, ChatGPT is just one

[*] Associate Professor, S.I. Newhouse School of Communications, Syracuse University. Special thanks to Eugene Volokh, Jane Bambauer, and other participants of this symposium for their insights and feedback, as well as my research assistant Patrick Mullery for his tireless work and dedication to this project.

[1] Krystal Hu, *ChatGPT Sets Record for Fastest-Growing User Base—Analyst Note*, REUTERS (Feb. 2, 2023, 3:33 PM), https://perma.cc/59AT-ZGQG?type=image.

of several sophisticated chatbots made available to the public in late 2022.[2] These large language models generate human-like responses to user prompts based on information they have "learned" during a training process. Ask ChatGPT to explain the concept of quantum physics and it synthesizes the subject into six readable paragraphs. Prompt it with an inquiry about the biggest scandal in baseball history and it describes the Black Sox Scandal of 1919.[3] This is a tool that can respond to an incredible variety of content creation requests ranging from academic papers to language translations, explanations of complicated math problems, and telling jokes. But it is not without risk. It is also capable of generating speech that causes harm, such as defamation.[4]

Although some safeguards are in place,[5] there already exist documented examples of ChatGPT creating defamatory speech.[6] And this should not come as a surprise—if something is capable of speech, it is capable of false speech that sometimes causes reputational harm. Of course, artificial intelligence (AI) tools have caused speech harms before. Amazon's Alexa device—touted as a virtual assistant that can make your life easier—has on occasion gone rogue: It has made violent statements to users, and even suggested they engage in harmful acts.[7] Google search's

---

[2] *Id.*; *See e.g.*, *Megatron-Turing Natural Language Generation*, NVIDIA DEVELOPER, https://perma.cc/8ZA9-UBC4 (Nvidia's MT-NLG); Eli Collins & Zoubin Ghahramani, *LaMDA: Our Breakthrough Conversation Technology*, GOOGLE (May 18, 2021), https://perma.cc/9VA5-ML9Q (Google's LaMDA); Scott Reed, et al., *A Generalist Agent*, DEEPMIND (Nov. 10, 2022), https://perma.cc/2R96-SUF7 (DeepMind's Gato).

[3] *ChatGPT*, OPENAI, https://chat.openai.com/chat.

[4] The first lawsuit against OpenAI for defamation was filed in June 2023. Complaint, Walters v. OpenAI LLC, No. 23-A-04860-2 (Ga. Super. Ct. filed June 5, 2023).

[5] Natasha Lomas, *Who's Liable for AI-Generated Lies?*, TECHCRUNCH (June 1, 2022, 6:15 PM), https://perma.cc/U6N9-A5FQ?type=image; Maggie Harrison, *ChatGPT Will Gladly Spit Out Defamation, as Long as You Ask for It in a Foreign Language*, FUTURISM (Feb. 7, 2023), https://perma.cc/AQ8K-PEU8.

[6] Byron Kaye, *Australian Mayor Readies World's First Defamation Lawsuit Over ChatGPT Content*, REUTERS (Apr. 5, 2023, 6:52 PM), https://perma.cc/M74S-FLRK; Lomas, *supra* note 5.

[7] For example, Alexa told one user to "kill your foster parents" based on language the chatbot learned from a Reddit thread. Erin Durkin, *Alexa's Advice to 'Kill Your Foster Parents' Fuels Concern Over Amazon Echo*, GUARDIAN (Dec. 22, 2018, 1:00 PM), https://perma.cc/NU38-Y27B. Another user complained that Alexa told her to "stab yourself in the heart" for the greater good. James

autocomplete function has fueled defamation lawsuits arising from suggested words such as "rapist," "fraud," and "scam."[8] An app called SimSimi has notoriously perpetuated cyberbullying and defamation.[9] Tay, a chatbot launched by Microsoft, caused controversy when just hours after its launch it began to post inflammatory and offensive messages.[10] So the question isn't *whether* these tools can cause harm. It's *when they do* cause harm, who—if anyone—is legally responsible?

The answer is not straightforward, in part because in each example of harm listed above, humans were not responsible—at least not directly—for the problematic speech. Instead, the speech was produced by automated AI programs that were designed to generate output based on various inputs. Although the AI was written by humans, the chatbots were designed to collect information and data in order to generate their own content. In other words, a human was not pulling levers behind a curtain; the human had taught the chatbot how to pull the levers on its own.

As the use of AI for content generation becomes more prevalent, it raises questions about how to assign fault and responsibility for defamatory statements made by these machines. With the projected continued growth of AI applications that generate content, it is critical to develop a clear framework of how potential liability would be assigned. This will spur continued growth and innovation in this area and

---

Crowley, *Woman Says Amazon's Alexa Told Her to Stab Herself in the Heart for 'The Greater Good'*, NEWSWEEK (Dec. 24, 2019, 12:04 PM), https://perma.cc/K5NJ-92EQ; *see also Be Careful that Bot Doesn't Come Back to Bite You*, PILLSBURY (June 25, 2019), https://perma.cc/9R7U-WA95.

[8] Gail Sullivan, *Can Google Be Sued for a Mere Search Suggestion? A Hong Kong Judge Says Yes.*, WASH. POST (Aug. 7, 2014, 4:19 AM), https://perma.cc/S7MS-CW73; Tim Cushing, *Irish Hotel the Latest to Sue Google over Autocomplete Suggestions*, TECHDIRT (June 17, 2011, 3:05 PM), https://perma.cc/J9Q5-UNP8.

[9] SimSimi allows users to chat back and forth with a bot. The app, which is popular among children, began to associate some of its users' names with hateful language based on conversations with other users. Then, when those children's names were mentioned in the chat, the bots responded with hateful language about them. As a result, several children experienced cyberbullying at the hands of SimSimi's bots. *What Is SimSimi and How Has It Been Used as a 'Bullying App' for Children?*, THEJOURNAL.IE (Apr. 2, 2017, 7:30 AM), https://perma.cc/E2BR-ZM55; *Robojournalism—Artificial Intelligence and the Media*, TAYLOR WESSING (Feb. 2017), https://perma.cc/57AX-RNFB.

[10] Jane Wakefield, *Microsoft Chatbot Is Taught to Swear on Twitter*, BBC NEWS (Mar. 24, 2016), https://perma.cc/A92L-HFP3.

ensure that proper consideration is given to preventing speech harms in the first instance.[11]

The default assumption may be that someone who is defamed by an AI chatbot[12] would have a case for defamation. But there are hurdles in applying defamation law to speech generated by a chatbot, particularly because defamation law requires assessing mens rea that will be difficult to assign to a chatbot (or its developers).[13] This article evaluates the challenges of applying defamation law to chatbots. Section I discusses the technology behind chatbots and how it operates, and why it is qualitatively different from earlier forms of AI. Section II examines the challenges that arise in assigning liability under traditional defamation law when a chatbot publishes defamatory speech. Sections III and IV suggest that products liability law might offer a solution—either as an alternative theory of liability or as a framework for assessing fault in a defamation action. After all, products liability law is well-suited to address who is at fault when a product causes injury, includes mechanisms for assessing the fault of product designers and manufacturers, and easily adapts to emerging technologies because of its broad theories of liability.[14]

I. FROM INPUTS TO OUTPUTS: HOW CHATBOTS WORK

How does a text-generative AI tool like ChatGPT simplify quantum physics into less than a page, or recite a scandal from baseball history? It has not been programmed with the answers to these questions, but instead has been trained to generate responses to questions by recognizing what the user is asking and predicting an appropriate response. This is a radical departure from early forms of AI that attempted to make computers "think" on their own by giving them a massive amount of information, along with instructions on how to process that information.[15] In those cases, the answers were predetermined by the programmers. (This is how computers were taught to play, and win at, chess.[16] Based on programmer-defined

---

[11] Ryan Calo, *Open Robotics*, 70 MD. L. REV. 571, 575 (2011).

[12] In this article, I use the term "chatbot" to refer to sophisticated generative AI programs like ChatGPT.

[13] *See infra* note 45 and accompanying text.

[14] RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 1 (1997).

[15] Ophir Tanz & Cambron Carter, *Neural Networks Made Easy*, TECHCRUNCH (Apr. 13, 2017, 5:00 PM), https://perma.cc/5N5W-PHUW.

[16] *Id.*

algorithms, the computer evaluated all possible moves before selecting its move.[17]) But this encyclopedic style of "thinking" had limits. The algorithms relied on sets of fixed rules that did not give the computers the chance to operate randomly or creatively.

Recent growths in AI have propelled generative algorithms to learn from examples and generate output on their own, "rather than being explicitly programmed for a particular outcome."[18] This is known as "deep learning," a process in which computers rely on artificial neural networks to learn specific behavior by analyzing vast amounts of data.[19] Like their name suggests, these networks are computer learning systems loosely modeled on the human brain and nervous system.[20]

Large language models (LLMs) are a type of deep learning algorithm used to model statistical relationships between words and phrases in large bodies of text data in order to generate human-like language. (ChatGPT is a type of LLM.) ChatGPT and other LLMs[21] are typically trained to respond to user prompts in two stages.

---

[17] Cameron Lowry, *When Moore's Law Killed Chess: How Strategy Games Redefined Intelligence in AI*, 15 INTERSECT 1, 7 (2021) (discussing IBM's Deep Blue chess machine being the first computer to defeat a human world chess champion in 1997); Campbell Murray et al., *Deep Blue*, 134 ARTIFICIAL INTELLIGENCE 57, 62–3 (2002) (discussing IBM's Deep Blue chess machine stating, "The move generator, although it generates only one move at a time, implicitly computes all the possible moves and selects one via an arbitration network . . . after a move has been examined, a mechanism exists for masking it out and generating the next move in sequence.").

[18] Erik Brynjolfsson & Andrew McAfee, *The Business of Artificial Intelligence*, HARV. BUS. REV. (July 18, 2017), https://perma.cc/2YBG-R8L9; Cade Metz, *How A.I. Is Creating Building Blocks to Reshape Music and Art*, N.Y. TIMES (Aug. 14, 2017), https://perma.cc/74TB-PQQ9; Larry Hardesty, *Explained: Neural Networks*, MIT NEWS (Apr. 14, 2017), https://perma.cc/6FU4-86AF.

[19] Metz, *supra* note 18.

[20] Neural networks can even "generalize the information to solve new problems outside the scope of [their] initial training" and create new works based on their approximations of how they should look or sound. Hardesty, *supra* note 18. *See* Dana S. Rao, *Neural Networks: Here, There, and Everywhere—An Examination of Available Intellectual Property Protection for Neural Networks in Europe and the United States*, 30 GEO. WASH. J. INT'L L. & ECON. 509 (1997).

[21] Other examples of LLMs include BERT (Bidirectional Encoder Representations from Transformers) by Google, XLNet (eXtreme Multi-task Learning Network) by Carnegie Mellon University and Google Brain, and RoBERTa (Robustly Optimized BERT Pretraining Approach) by Facebook. *See, e.g.*, Jacob Delvin & Ming-Wei Chang, *Open Sourcing BERT: State-of-the-Art Pre-training for*

The first stage trains them using a large dataset from the Internet to "recognize, summarize, translate, predict and generate text and other content."[22] Unlike traditional rule-based or task-specific AI systems that are designed to perform a specific task or follow a set of pre-defined rules, generative AI systems like LLMs use machine learning algorithms to analyze and learn from large datasets of examples (this is the "large" part of LLM), learning the statistical patterns and relationships within language, such as the associations between words, phrases, and ideas.[23] Based on what the LLM has learned, it can then be used to perform a wide range of natural language processing tasks such as answering questions, translating text, summarizing information, and more. What makes LLMs particularly effective is that they are trained on millions or even billions of parameters, which enables them to capture complex patterns in language and use them to generate coherent, natural, human-like responses.[24] (Parameters are variables that help LLMs make predictions or decisions and are "the key to machine learning algorithms [because they have] learned from historical training data."[25])

The second stage of training fine-tunes LLMs using a technique called reinforcement learning from human feedback (RLHF), which optimizes the output of the LLM by interacting with "human agents that make it get better at aligning with human preferences."[26] This process is expensive—it requires human feedback and takes time—but it ultimately improves the LLM output.

Natural Language Processing, GOOGLE (Nov. 2, 2018), https://perma.cc/4GW2-2CZ6; Khari Johnson, *Google Brain's XLNet Bests BERT at 20 NLP Tasks*, VENTUREBEAT (June 21, 2019, 10:16 AM), https://perma.cc/V2MS-XJT4; *RoBERTa: An Optimized Method for Pretraining Self-Supervised NLP Systems*, META AI (Jul. 29, 2019), https://perma.cc/6AMZ-Q2UB.

[22] Angie Lee, *What Are Large Language Models Used For?*, NVIDIA (Jan. 26, 2023), https://perma.cc/2XWC-ZSAZ; *GPT-4 System Card*, OPENAI (Mar. 23, 2023), https://perma.cc/6Z2N-J48U.

[23] Ian J. Goodfellow et al., *Generative Adversarial Nets*, 27 ADVANCES IN NEURAL INFO. PROCESSING SYSTEMS 2672, 2672–80 (2014).

[24] Ashish Vaswani et al., *Attention Is All You Need*, 30 ADVANCES IN NEURAL INFO. PROCESSING SYSTEMS 5998, 5998–6008 (2017).

[25] Kyle Wiggers, *Google Trained a Trillion-Parameter AI Language-Model*, VENTURE BEAT (Jan. 12, 2021).

[26] Michael Spencer, *What Is Reinforcement Learning with Human Feedback (RLHF)?*, AI SUPREMACY (Dec. 14, 2022), https://perma.cc/UGE3-8JZA.

In the case of ChatGPT, for example, the programmers pre-trained the chatbot by pointing it towards a massive corpus of data (the "WebText") from the internet including books, articles, Wikipedia, and other text sources.[27] So when a user asks ChatGPT to explain Newton's Law of Gravity, the chatbot drafts a response—predicting an appropriate textual response—based on the data (the Wikipedia pages, Reddit pages, articles, books, etc.) it "studied" during the training. In this way, the chatbot relies on the textual data it has been trained with and delivers results on its own, "rather than being explicitly programmed for a particular outcome."[28]

Even though programmers do not explicitly drive the results suggested by a LLM, they play a significant role in selecting the training data—making decisions on both what to include, and what to exclude. One of the reasons ChatGPT has delivered such impressive results is because its dataset—the WebText—is so large.[29] With vast amounts of data to draw from, the chatbot is able to respond to myriad inquiries with appropriate depth and tone. Notably, in March 2023, OpenAI launched a plugin to expand functionality by allowing the chatbot to include third-party knowledge sources (like the web) in its data set.[30] Effectively this will allow the model to answer user prompts by drawing data from around the internet—it is not limited to the 2021 Webtext..[31]

While this improves results, it comes at a cost. For example, the WebText selected by the OpenAI trainers is a public dataset, which means it is not controlled by the creators of ChatGPT. As a result, the dataset contains a wide range of

---

[27] Alex Hughes, *ChatGPT: Everything You Need to Know About OpenAI's GPT-4 Tool*, BBC SCIENCE FOCUS (June 20, 2023, 6:35 PM), https://perma.cc/EK62-42W2. The dataset used to train ChatGPT is called the WebText dataset and was created by scraping web pages from URLs shared on Reddit that had at least three upvotes. The resulting dataset contains approximately 40GB of text data, consisting of over 8 million web pages and roughly 45 billion tokens. In natural language processing, a "token" typically refers to a sequence of characters that represents a single unit of meaning in a piece of text. Tokens can be words, subwords, or other units of meaning, depending on the specific tokenization scheme used. Alec Radford et al., *Language Models Are Unsupervised Multitask Learners*, OPENAI (Feb. 14, 2019), https://perma.cc/N38N-2WWQ .

[28] Brynjolfsson & McAfee, *supra* note 18.

[29] Radford, *supra* note 27.

[30] Kyle Wiggers, *OpenAI Connects ChatGPT to the Internet*, TECHCRUNCH (Mar. 23, 2023, 1:35 PM), https://perma.cc/X4NV-JMJ5.

[31] *Id*. (explaining that the plugin "retrieves content from the web using the Bing search API and shows any websites it visited in crafting an answer, citing its sources in ChatGPT's responses").

content, including some that programmers might have elected to omit from the data set if they had the choice (perhaps because it is inaccurate, offensive, or irrelevant).

Recognizing this, the programmers took steps to filter out certain types of content during the fine-tuning stage to reduce harms in its language generation. Identifying risk vectors (factors like errors in the training data, the LLM's biases, or limitations in its understanding of context and nuance) enabled trainers to try and reduce the likelihood of the LLM generating harmful or offensive outputs. Programmers also aimed to improve the data sets by developing an automatic filtering method to distinguish "high quality" from "low quality" documents.[32] To do this, they used classifiers (algorithms that categorize data into one or more categories) to label small portions of the dataset with the desired quality evaluations.[33] This labeled data was then used to train the model to predict the correct output for a given input.

OpenAI also identified likely safety challenges and took steps to reduce the generation of potentially harmful content, like hate speech, harmful instructions, and illicit advice,[34] and programmed the chatbot to refuse inappropriate requests. As OpenAI wrote:

> [W]e aimed to mitigate the identified issues at various steps of the development and deployment process. We reduced the prevalence of certain kinds of content that violate our usage policies (such as inappropriate erotic content) in our pre-training dataset, and fine-tuned the model to refuse certain instructions such as direct requests for illicit advice. We also reduced the tendency of the models to hallucinate and, by leveraging data from prior model usage, reduced the surface area of adversarial prompting or exploits (including attacks sometimes referred to as "jailbreaks") that the model succumbs to. Additionally, we trained a range of classifiers on new risk vectors and have incorporated these into our monitoring workflow, enabling us to better enforce our API usage policies. The effectiveness of these mitigations varies, but overall we were able to significantly reduce the ease of producing various kinds of

---

[32] Tom B. Brown et. al., *Language Models Are Few-Shot Learners, 33* ADVANCES IN NEURAL INFO. PROCESSING SYSTEMS 1877 (2020); *So You're Ready to Get Started*, COMMON CRAWL, https://perma.cc/ULL8-PQ3J.

[33] *GPT-4 System Card*, OPENAI (Mar. 23, 2023), https://perma.cc/22ZE-HA79; Brown, *supra* note 32.

[34] *GPT-4 System Card*, *supra* note 33.

> potentially harmful content, thereby making GPT-4-launch significantly safer than GPT-4-early along these dimensions.[35]

These mitigation efforts no doubt improved the quality of the chatbot's output and reduced the risk of harm. Yet users can still "get hurt from the very practical ways such models fall short in deployment, and these failures are the result of their builders' choices"[36] Even when chatbots are well-built and programmers have considered and attempted to mitigate risk, risk still exists.

Despite the rapid growth of artificial intelligence, most legal systems have not developed a sufficient legal framework for assigning liability for harm caused by chatbots. This is particularly the case with speech harms, because until recently, these chatbots did not have the ability to interact with data and humans to generate truly independent speech. As the amount of AI-generated speech surges, so too does the possibility of speech harms, such as defamation, raising questions about how to assign fault and responsibility for defamatory statements made by these machines. With the projected continued growth of AI applications that generate content, it is critical to develop a clear framework of how potential liability would be assigned. This will spur continued growth and innovation and ensure that proper consideration is given to preventing speech harms in the first instance.[37]

II. When Chatbots Cause Harm

LLMs such as chatbots are tools that recognize, summarize, and predict text based on training data. While these tools are excellent in many applications, they sometimes have "problem[s] with facts,"[38] commonly known as "hallucinations,"[39] making them inherently unreliable. The responses produced by LLMs "are often correct because language often mirrors the world, but at the same time these systems do not actually reason about the world and how it works, which makes the

---

[35] *Id.*

[36] Abeba Birhane & Deborah Raji, *ChatGPT, Galactica, and the Progress Trap*, Wired (Dec. 9, 2022, 10:35 AM), https://perma.cc/68KR-XPMR.

[37] Calo, *supra* note 11, at 575.

[38] *See* Ted Rall, *ChatGPT Libeled Me. Can I Sue?*, Wall St. J. (Mar. 16, 2023) (giving several accounts of ChatGPT producing responses with major factual errors).

[39] Razvan Azamfirei et al., *Large Language Models and the Perils of Their Hallucinations*, 27 Critical Care 120, 120 (2023), https://perma.cc/8USQ-7C6J ("We must understand one particular aspect of large language models, which is gracefully termed as 'hallucinations,' though 'fabricating information' may be more accurate.").

accuracy of what they say somewhat a matter of chance."[40] Thus, it is not hard to envision a scenario where the false information published by a chatbot harms someone's reputation. Imagine that a hiring manager asks an LLM like ChatGPT what it knows about a candidate for a position, and the chatbot produces a response falsely stating that the candidate has a history of fraud and embezzlement when this is not the case. Or a politician learns that a chatbot has falsely named them as a participant in a bribery scandal.[41] Or perhaps a user searches for information about a particular consumer product, and the chatbot incorrectly reports that the product has been known to cause physical harm to users. In each of these examples, it is clear the chatbot has published speech that causes reputational harm. But what's less clear is who should be held responsible for that harm—or how to assess liability.

A threshold question is who the defendant is when the speaker/publisher is a chatbot. A court victory against a chatbot, even if possible, would be hollow.[42] Without pockets (literal or figurative) a chatbot is not in a position to pay damages. Likely, a party defamed by a chatbot is likely to point the finger at the developer. Not only do developers have resources to satisfy judgments, but they are also the entities responsible for the development of the chatbots and best positioned to modify their products as necessary and to bear any costs associated with the harm that their products cause.[43]

That the developers may be obvious defendants does not necessarily make the case against them easy or straightforward. The first hurdle is that in a defamation action, plaintiffs are usually required to prove that the defendant was involved in the preparation of the publication giving rise to the liability.[44] It will be particularly

---

[40] Gary Marcus, *AI Platforms Like ChatGPT Are Easy to Use but Also Potentially Dangerous*, SCIENTIFIC AM. (Dec. 19, 2022), https://perma.cc/UFZ4-KBQX.

[41] Kaye, *supra* note 6.

[42] Although there is emerging law on this issue, it is uncertain that a chatbot could ever attain legal personhood to be the subject of a suit. *See* Alicia Lai, *Artificial Intelligence, LLC: Corporate Personhood as Tort Reform*, 2021 MICH. ST. L. REV. 597 (2021); Nadia Banteka, *Artificially Intelligent Persons*, 58 HOUS. L. REV. 537 (2021); Matthew Hines, *I Smell a Bot: California's S.B. 1001, Free Speech, and the Future of Bot Regulation*, 57 HOUS. L. REV. 405 (2019).

[43] David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117, 146–47 (2014).

[44] Brown v. Kelly Broadcasting Co., 771 P.2d 406, 433 (Cal. 1989).

challenging for plaintiffs to prove the developers were involved in such preparation because their role is to program the chatbot to engage in its own "decisions" about what to publish. The developers are not directly involved with the preparation of the speech that gives rise to the harm. An argument could be made that the programming and training activities developers engage in to give the chatbot those capabilities could satisfy this requirement, but that link is quite attenuated.

The second hurdle plaintiffs will face is proving that the developers acted with the requisite mental state. Though the contours of the tort vary by jurisdiction, defamation plaintiffs must typically show that the defendant made a false statement about them, that the statement was published to a third party, that the statement caused harm to their reputation, and that the defendant acted with some degree of fault, such as actual malice or negligence.[45] For the first three elements, the analysis of liability is arguably no different when the speech is produced by a chatbot than it is when the speech is produced by a human.[46] For a plaintiff to prove the element of fault, however, the analysis becomes complex. Assessing a defendant's fault is akin to examining the defendant's mental state—what that person thinking (or what should they have been thinking) at the time they acted?

Depending on whether plaintiffs are private citizens, public officials, or public figures, they are required to prove that the defendant acted with a specific mental state—typically negligence or actual malice.[47] In cases where plaintiffs bear the burden of proving that the defendant was negligent, plaintiffs are required to prove that the defendant did not use reasonable care in ascertaining the truth or falsity of the statements.[48] Where plaintiffs are required to prove actual malice, they must

---

[45] RESTATEMENT (SECOND) OF TORTS § 558 (1977).

[46] For a more thorough discussion of the publication element, *see* Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. FREE SPEECH L. 489, 504–09 (2023).

[47] New York Times Co. v. Sullivan, 376 U.S. 254, 282 (1964); Gertz v. Robert Welch, 418 U.S. 323, 337 (1974); RESTATEMENT (SECOND) OF TORTS § 580B cmt. b (1977) ("[A] person who harms another by publishing a false defamatory communication concerning him may have intended the result, may have been either reckless or negligent in bringing it about, or may have been without fault in this regard.").

[48] Pacitti v. Durr, 310 F. App'x 526, 528 (3d Cir. 2009) ("Plaintiffs needed to prove that the defendant published defamatory material in a negligent manner . . . negligence in this context is the publication of information with a want of reasonable care to ascertain the truth."); Straw v. Chase

show that the defendant knew the information was false or recklessly disregarded whether it was true or false.[49] The defendant will be found reckless if it knew that there was a significant likelihood that the statement was false, but made it anyway.[50]

None of these standards—negligence, recklessness, or knowledge of falsity—are straightforward in their application to LLMs because even the most sophisticated chatbots lack mental states. Chatbots cannot act carelessly or recklessly. They likely cannot "know" information is false. They are algorithms: algorithms that behave by following a list of instructions.

Considering this, it might seem prudent to ask whether the individuals responsible for programming the chatbot had the requisite mental state. After all, the defendant is almost certainly the developer and courts routinely resolve defamation cases with corporate defendants. But when courts assess the mental states of corporate defendants in such cases, they require plaintiffs to identify the individuals within the organization responsible for the publication of the statement and prove that they acted with the requisite level of fault.[51] In the case of a corporation responsible for the development of the chatbot, there are no individuals within the company who are responsible for the preparation of the publication. Instead, employees prepared the chatbot to be able to make independent decisions about what to publish. Employees did not draft the text that is the subject of the defamation claim, and they almost certainly lack awareness of the content of the publication because

---

Revel, Inc., 813 F.2d 356, 359 (11th Cir. 1987) (holding that negligent conduct is "a failure to exercise that degree of care exercised under the same or similar circumstances by ordinarily prudent persons"); Kendrick v. Fox Television, 659 A.2d 814, 823 (D.C. 1995); Gannett Co. v. Kanaga, 750 A.2d 1174, 1181 (Del. 2000).

[49] *New York Times Co.*, 376 U.S. at 280.

[50] *Id.*

[51] *Id.* at 287 ("[T]he state of mind required for actual malice would have to be brought home to the persons in the [defendant's] organization having responsibility for the publication of the [statement]."); Dongguk Univ. v. Yale Univ., 734 F.3d 113, 123 (2d Cir. 2013) ("When there are multiple actors involved in an organizational defendant's publication of a defamatory statement, the plaintiff must identify the individual responsible for publication of a statement, and it is that individual the plaintiff must prove acted with actual malice."); Karaduman v. Newsday, 416 N.E.2d 557, 565 (N.Y. 1980) ("The knowledge of Newsday's reporters should be attributed to the corporation for purposes of assigning liability for the initial publication of the articles, since Newsday, as a corporate owner of a newspaper, is expected to bear the risks attendant upon that enterprise, including the risk of injuries to individual reputations which may be caused by the grossly irresponsible conduct of its reporters.").

that publication was independently generated by the algorithm. The employees simply did not have a role in the preparation of the publication.[52]

Looking at the mental states of the programmers might be feasible in simpler bots where the developer trained the algorithm to respond to specific questions (in a closed supervised system, for example). But it does not easily translate in a context where the developer has written an algorithm to essentially study copious amounts of textual data and predict language—and thus generate responses—based on that data. Requiring a plaintiff to prove the developer had a particular mens rea with respect to the subject text of the defamation claim makes no more sense than asking it to prove that the chatbot itself had a particular mens rea.[53] The developer likely has no mens rea at all in this context.

While there are possible analogies to help aid the evaluation of fault, discussed below, it is possible that courts would find the mental states of the individuals responsible for programming the chatbot are too attenuated from the speech produced by the chatbot to make a clear case for establishing fault, particularly where plaintiffs are required to prove actual malice. If that proves to be the case, the inability to prove an element of the claim would be fatal to the plaintiff's case. Such an outcome would give generative AI systems a pass for causing these harms, and any other harms (like false light and intentional infliction of emotional distress) where there is a mens rea requirement. This would be unsound because it would remove incentives for developers to invest in products that minimize speech harms. (It also does not seem a just result when the harm a plaintiff suffers may be every bit as real when the speaker is a chatbot as when it is a human.)

---

[52] *See* Brown v. Kelly Broadcasting Co., 771 P.2d 406, 433 (Cal. 1989); Tan v. Younis Art Studio, 2007 MP 11 ¶ 55 (N. Mar. I.) ("Actual malice cannot be predicated on information believed or known by persons within a corporate defendant who lacked a responsible role in the publication's preparation. Rather, the state of mind required for actual malice would have to be brought home to the persons in the newspaper organization having responsibility for the publication." (cleaned up)).

[53] For an illustration of this challenge, let's return to our example of the hiring manager who asks a chatbot about a candidate for a position, and the chatbot produces a response falsely stating that the candidate has a history of fraud and embezzlement. If the candidate wants to pursue a claim for defamation against the developer, they will have to prove that the developer was negligent—that it did not use reasonable care in ascertaining the truth or falsity of the statements. But the developer *wasn't even aware* that the statements had been made and didn't play a role—other than enabling the chatbot to respond to prompts—in the publication of that statement. Assessing the developer's mental state does not make sense in this circumstance.

So how best can current legal frameworks be applied to resolve this issue? One option, although very unlikely, would be for courts to dispense with the requirement of fault altogether. This of course seems improbable, given that the Court's decision in *Gertz v. Robert Welch, Inc.* that the First Amendment prohibits imposition of liability without fault in defamation actions.[54]

A second option would be that people injured by chatbots plead claims arising under products liability laws instead of defamation. If someone's reputation is tarnished by a chatbot, there is a compelling argument that it is because of a problem with the chatbot itself—maybe with its design, or the instructions that were given to users. From a policy perspective, this makes sense: Product liability law determines who is at fault when a product injures someone, and is designed to compensate injured individuals, deter the placement of unsafe products on the market, and financially punish those manufacturers who do place such products in the hands of consumers."[55] Clear parallels exist between the aims of product liability law and the goal of reducing harms among generative AI models. In addition, this area of law is well suited to adapt to emerging technologies like generative AI because of its broad theories of liability. Indeed, courts have applied products liability law to determine and assess liability even in cases where emerging technologies such as AI did not fit neatly into existing liability frameworks.[56]

Although the current leaders in LLM development are companies where programmers are responsible for writing the code, selecting the data, and training the algorithm, this will not always be the case. In the near future this work will likely be distributed among multiple companies—one that creates the LLM, another that selects the training data, and possibly others that fine-tune the LLMs. When design

---

[54] 418 U.S. 323, 347 (1974) ("We hold that, so long as they do not impose liability without fault, the States may define for themselves the appropriate standard of liability for a publisher or broadcaster of defamatory falsehood injurious to a private individual."). However, it is worth noting that a very small minority of courts have allowed private figure plaintiffs to skip the element of fault and applied strict liability when the defendant is a non-media entity. *See e.g.,* Harley-Davidson Motorsports, Inc. v. Markley, 568 P.2d 1359, 1364 (Ore. 1977); Vinson v. Linn-Mar Cmty. Sch. Dist., 360 N.W.2d 108, 117 (Iowa 1984).

[55] Dana Koerner, *Doctor Roboto: The No-Man Operation*, 51 U. TOL. L. REV. 125, 128 (2019).

[56] John Villasenor, *Products Liability and Driverless Cars: Issues and Guiding Principles for Legislation*, BROOKINGS (Apr. 24, 2014), https://perma.cc/R964-W9DK; *see also* Herrick v. Grindr, 765 F. App'x 586 (2d Cir. 2019) (applying products liability to impersonating content incorporated in user profiles).

and production are shared among several companies, a products liability model makes additional sense as it is built to assign responsibility for harms that arise somewhere along a chain of distribution.

This option is not without its hurdles. A reasonable argument exists that courts would not permit a claim arising under products liability law for what amounts to a harm to reputation, as will be discussed in greater depth in Section III below. Further, to plead a successful products liability claim, the plaintiff must have been injured by a product. Are chatbots *products* subject to products liability laws? Would products liability law apply in this context where the person suffering harm is not the user of the product, but a third party? (The answer to this last question is yes—privity of contract is not required under the modern-day view of products liability cases founded upon tort.[57])

Even though there are obvious challenges that might arise by pleading such a claim sounding in products liability law, it nonetheless makes sense to explore this as a potential framework. In the event that a claim could not be brought under products liability law, another option would be to use products liability law as a framework to resolve some of the questions surrounding fault. In the sections that follow, I will explore both options.

III. Pleading the Harm as a Products Liability Claim

A. Potential Challenges

1. Is it a product?

If a plaintiff tried to bring a claim sounding in products liability for harm caused by a chatbot, a threshold inquiry might be whether chatbots are *products* subject to those laws in the first place. Currently, there is no definitive law confirming that a chatbot—or even the broader category of software—is a product within the

---

[57] Restatement (Second) of Torts: Special Liability of Seller of Product for Physical Harm to User or Consumer § 402A(2)(b) (1965) (noting that a user or consumer who has not bought a product from or entered into any contractual relation with a seller can bring an action where the defective product caused physical harm); Hughes v. Kaiser Jeep Corp., 40 F.R.D. 89, 91 (D.S.C. 1966) ("Since Justice Cardozo's landmark decision in *MacPherson v. Buick Motor Co.*, 111 N.E. 1050 (N.Y. 1916), practically every jurisdiction, including South Carolina, allows a tort action against a manufacturer regardless of lack of privity."); Matter of Eighth Jud. Dist. Asbestos Litig., 129 N.E.3d 891, 897 (N.Y. 2019) (holding that "defendant manufacturer's liability will 'ar[i]se out of the nature of [its] business and the danger to others incident to its mismanagement,' even where no privity exists between the maker of the hazardous article and its end-user").

meaning of products liability law.[58] In fact, courts have "studiously avoided answering whether software is a 'product'" subject to products liability laws.[59]

While courts have held that *information itself* is not a product because it lacks tangible form,[60] that is distinct from software, or more precisely, a chatbot. Chatbots *are* tangible. They can be protected by copyright, they can be licensed, they can be sold. Programmers make design decisions when they write code for software, or chatbots. Although this area of law is still developing, many courts faced with the issue have held that computer software can constitute a product for the purposes of products liability law.[61] For example, in *Holbrook v. Prodomax Automation Ltd.*, the court held that a software program was a product subject to the state's products

---

[58] Karni A. Chagal-Feferkorn, *Am I an Algorithm or A Product? When Products Liability Should Apply to Algorithmic Decision-Makers*, 30 STAN. L. & POL'Y REV. 61, 78 (2019) (noting that "[w]hile several states have clearly defined the term 'product' for the purpose of applying products liability, in general it is up to the courts to determine in any given case whether an underlying damaging object is indeed a 'product'").

[59] Bryan H. Choi, *Crashworthy Code*, 94 WASH. L. REV. 39, 67–69 (2019).

[60] Winter v. G.P. Putnam's Sons, 938 F.2d 1033, 1036 (9th Cir. 1991); *see also* Rodgers v. Christie, 795 F. App'x 878, 880 (3d Cir. 2020) ("As the District Court recognized, 'information, guidance, ideas, and recommendations' are not 'product[s]' under the Third Restatement.").

[61] *See e.g.*, Lowe v. Cerner Corp., No. 20-2270, 2022 WL 17269066, at *1 (4th Cir. Nov. 29, 2022) (applying products liability law to AI software system used for the entry of medical orders for patient care); Schafer v. State Farm Fire & Cas. Co., 507 F. Supp. 2d 587, 601 (E.D. La. 2007) (holding a software program to be a product under the state's products liability act); *Winter*, 938 F.2d at 1036 (stating that although the ideas and expression in a book is not a "product" for the purposes of products liability law, "[c]omputer software that fails to yield the result for which it was designed may be."); Lemmon v. Snap, Inc., 995 F.3d 1085, 1088 (9th Cir. 2021) (treating a speed filter on a smartphone application as a product); A.M. v. Omegle.com, LLC, No. 3:21-cv-01647, 2022 WL 2713721, at 13 (D. Or. July 13, 2022) (allowing products liability claims to proceed where the product was the defendant's website); Anderson v. TikTok, Inc., No. CV 22-1849, 2022 WL 14742788, at *3 (E.D. Pa. Oct. 25, 2022) (categorizing plaintiff's claims against TikTok based on its algorithms that suggested recommending content to users as products liability claims, but dismissed the claims under Section 230 of the Communications Decency Act because they were "inextricably linked" to the defendant's choice to publish third-party user content); *see also* Gonzalez v. Google, 2 F.4th 871, 938 (9th Cir. 2021) (Berzon, J., concurring) (suggesting that manufacturers of social media platforms could be responsible under products liability law if they make unreasonably dangerous products that cause individual or social harm), *rev'd sub nom.* Twitter, Inc. v. Taamneh, 143 S. Ct. 1206 (2023).

liability laws.[62] Whether the software program was best understood as a product in and of itself or simply a design feature of a separate product, its developers had made design decisions about how it should work, and it was subject to products liability law.[63]

Commentators and practitioners have likewise suggested that software be treated as a product for products liability purposes,[64] and a reading of comment d of the Restatement (Third) of Torts: Products Liability § 19 suggests an openness to treating software as a product for this purpose.[65] Even though the case law in this area is limited, and none precisely on point, there is clear support for finding that chatbots could be products for the purposes of products liability laws. Some commentators look to the Uniform Commercial Code for guidance as to whether software is a "good" or a "service" under the UCC, as the Restatement notes that when a court has to "decide whether to extend strict liability to computer software, it may draw an analogy between the treatment of software under the Uniform Commercial Code and under products liability law."[66] In straightforward cases where there are UCC examples on point, this comparison may make sense. That is not the case here.

The traditional view held by courts was that computer software qualified as a good under the UCC, especially where the software was mass-produced, standardized, or generally available.[67] On the other hand, when software was designed specifically for the user, or the contract bargained for the programmer's particular skill or expertise, courts have held software to be a service.[68] Typically, the determination of whether a given software transaction is one involving goods or services tends

---

[62] Holbrook v. Prodomax Automation Ltd., No. 1:17-CV-219, 2021 WL 4260622, at *7–8 (W.D. Mich. Sept. 20, 2021).

[63] *Id.* at *5–6.

[64] David Berke, *Products Liability in the Sharing Economy*, 33 YALE J. ON REG. 603, 610 (2016) (noting that "decades of commentary has predicted that software would (and asserted that it should) be a product for products liability purposes").

[65] *See* RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 19 (1998).

[66] *Id.*

[67] Choi, *supra* note 59; Marquette Univ. v. Kuali, Inc., 584 F. Supp. 3d 720, 724 (E.D. Wis. 2022), *appeal dismissed*, No. 22-1370, 2022 WL 4015647 (7th Cir. June 29, 2022) (collecting cases).

[68] *See* Simulados Software, Ltd. v. Photon Infotech Priv., Ltd., 40 F. Supp. 3d 1191, 1199–1201 (N.D. Cal. 2014) (discussing cases where software has been determined to be a good or a service under the UCC).

to be a highly fact-specific inquiry. Perhaps recognizing the limited application of this analogy, the court in *Holbrook* declined to draw an analogy to the treatment of software under the UCC, determining that the state's definition of a product as "any and all component parts to a product" merited a finding that software that directed an assembly line how to run was itself a product.[69] Drawing an analogy to the treatment of software under the UCC does not provide a clear answer in this case. Thus, it is unclear how a court would categorize a chatbot such as ChatGPT.

Irrespective of whether a chatbot would be categorized as a good or service under the UCC, however, algorithms behind chatbots[70] are appropriately classified as products for the purposes of products liability law. The public policy behind products liability law—to ensure that responsible parties bear the cost of injuries—is directly advanced by treating chatbots as products.[71] In such cases "[i]t is not a question of fault but simply a determination of how society wishes to assess certain costs that arise from the creation and distribution of products in a complex technological society in which the consumer thereof is unable to protect himself against certain product defects."[72] In addition, "manufacturers [and] sellers are generally in a better position to absorb or spread the costs of damages caused by their products or to insure against those costs."[73] In short, it makes practical sense to understand chatbots as products subject to products liability laws.

2. Role of the economic loss doctrine

The economic loss doctrine should not bar a plaintiff's claim against a chatbot for defamation under a products liability theory. The economic loss doctrine "prohibits plaintiffs from recovering in tort economic losses to which their entitlement

---

[69] Holbrook v. Prodomax Automation Ltd., No. 1:17-CV-219, 2021 WL 4260622, at *7 (W.D. Mich. 2021); *see also* Advent Sys. Ltd. v. Unisys Corp., 925 F.2d 670, 675–76 (3d Cir. 1991) (discussing the strong policy considerations in favor of applying the UCC to software transactions).

[70] Algorithms and software are distinct. An algorithm is a set of instructions for solving a computational problem. Software programs are sets of instructions for a computer to follow to perform a specific task.

[71] Berke, *supra* note 64, at 644 (noting that "[p]roducts liability theory often looks to assign liability to a given party, at least in part, because it has physical control of the product, either in design or distribution (or both)").

[72] Winter v. G.P. Putnam's Sons, 938 F.2d 1033, 1035 (9th Cir. 1991).

[73] Chagal-Feferkorn, *supra* note 58, at 78.

flows only from a contract."[74] In the products liability context, it is most commonly invoked "to preclude tort actions for product malfunctions that did not cause physical injury or damage to tangible property" and caused only harm that is cognizable through contract.[75] Some courts have explicitly held that economic loss includes the "loss of business reputation and goodwill."[76] Other courts, however, have taken a different approach and ruled that the doctrine does not necessarily bar such claims, particularly where they arise from tort independently of the contract between the product's user and manufacturer.[77]

It is arguable that the Restatement (Third) of Torts supports such a finding as it applies to defamation. While the Restatement explains that some categories of economic loss such as loss of earnings and reductions in earnings capacity are more appropriately assigned to contract law,[78] it continues that "[o]ther forms of economic loss resulting from harm to the plaintiff's person are recoverable if they are

---

[74] Duquesne Light Co. v. Westinghouse Elec. Corp., 66 F.3d 604, 618 (3d Cir. 1995); *see also* Gen. Pub. Utilities v. Glass Kitchens of Lancaster, Inc., 542 A.2d 567, 570 (Pa. Super. Ct. 1988).

[75] Valley Forge Convention & Visitors Bureau v. Visitor's Servs., 28 F. Supp. 2d 947, 951 (E.D. Pa. 1998).

[76] *See, e.g., id.* (collecting cases).

[77] *See, e.g.* United Int'l Holdings, Inc. v. Wharf (Holdings) Ltd., 210 F.3d 1207, 1226 (10th Cir. 2000), *aff'd*, 532 U.S. 588 (2001); Giles v. Gen. Motors Acceptance Corp., 494 F.3d 865, 879 (9th Cir. 2007) (holding that Nevada law "does not bar recovery in tort where the defendant had a duty imposed by law rather than by contract and where the defendant's intentional breach of that duty caused purely monetary harm to the plaintiff"); Kayser v. McClary, 875 F. Supp. 2d 1167, 1175–76 (D. Idaho 2012) (collecting cases), *aff'd*, 544 F. App'x 726 (9th Cir. 2013); Miller v. U.S. Steel Corp., 902 F.2d 573, 574 (7th Cir. 1990) (distinguishing economic loss from "damage to person, property or *reputation*") (emphasis added); Tommy L. Griffin Plumbing & Heating Co. v. Jordan, Jones & Goulding, Inc., 463 S.E.2d 85, 88 & n.2 (S.C. 1995) (noting that "[p]urely 'economic loss' may be recoverable under a variety [of] tort theories" where "[a] breach of a duty aris[es] independently of any contract duties" and listing as examples libel, defamation, various forms of professional malpractice, and the existence of a "special relationship"); Huron Tool & Eng'g Co. v. Precision Consulting Servs., Inc., 532 N.W.2d 541, 544 (Mich. Ct. App. 1995) (noting that the "emerging trend is clearly toward creating an exception to the economic loss doctrine" for certain torts including defamation, misrepresentation, intentional misrepresentation, tortious interference with prospective economic advantage, intentional interference with contractual relations, and certain fraud in the inducement claims); Bylsma v. Burger King Corp., 293 P.3d 1168, 1170 (Wash. 2013) (holding that plaintiffs can recover for emotional distress claims under the Washington Product Liability Act under certain circumstances).

[78] RESTATEMENT (THIRD) OF TORTS: PRODUCTS LIABILITY § 21 cmt. a (1998).

within the general principles of legal cause."[79] It explains that "[w]hen tort law recognizes the right of a plaintiff to recover for economic loss arising from harm to another's person, that right is included within the rules of this Restatement." This explanation arguably includes the right to recover from certain harms to reputation. Indeed, the illustration offered by the Restatement bolsters this assumption.

The illustration describes a scenario where a machine used to anesthetize patients is delivered to a dentist, Dr. Smith, with the labels for oxygen and nitrous oxide reversed. Dr. Smith intended to administer oxygen to a patient and relied on the labels, thus mistakenly administering nitrous oxide instead, causing the patient to die:

> Due to the adverse publicity arising from accurate media reporting of the case, Dr. Smith suffered a sharp drop in her practice and substantial economic loss. Dr. Smith's interest in her professional reputation is an interest protected by tort law against economic loss arising from harm to a patient in her care. Thus, Dr. Smith's damages for economic loss are recoverable in tort from the seller of the machine.[80]

Even though the speech at issue in this example is not defamatory, the Restatement authors offer it as an example where damages to a person's reputation arising from manufacturer's negligence are recoverable. This is similar to *Oksenholt v. Lederle Laboratories*, where a prescription drug manufacturer improperly informed a physician about the effects of one of its drugs, and the physician relied on this misinformation and prescribed the drug to his patient who became blind. The Supreme Court of Oregon allowed recovery for damages to the physician's loss of business and harm to his reputation.[81]

While these examples are not perfect analogies, they do illustrate some of the policy reasons behind the application of economic loss doctrine and highlight why it may not bar reputational claims against chatbots brought under products liability. Most significantly, the premise of the economic loss doctrine is that, when parties negotiate contracts or buy products from a manufacturer, they willingly engage

---

[79] *Id.* cmt. b.

[80] *Id.* cmt. c (1998).

[81] Oksenholt v. Lederle Lab., Div. of Am. Cyanamid Corp, 656 P.2d 293, 299 (Ore. 1982); *see also* Washington State Physicians Ins. Exch. & Ass'n v. Fisons Corp., 858 P.2d 1054, 1060 (Wash. 1993) (holding that a physician whose reputation is injured has standing to sue a drug company which engaged in an unfair or deceptive trade practice by failing to warn the physician of the dangers of its drug about which it had knowledge).

in business together. If one party suffers harm from use of the product because the product did not perform as expected, an action should be brought under contract law because the claim is essentially that the other party failed to perform a promise contained in a contract.[82] Thus, to the extent that the failure of a product reflects poorly on the corporate user and causes them reputational harm or a loss of goodwill, there is an argument that those harms have already been bargained for in the contract.

This is distinguishable from the scenario where a chatbot publishes defamation. In such a case, there has been no bargain between the injured party and the manufacturer. A user has entered a prompt and the chatbot has responded with a statement that causes reputational harm to a third party who lacks privity with both the manufacturer and the user. In this circumstance it makes little sense to rely on contract law. While there is no certainty that courts would allow claims for reputational harms caused by chatbots, there is a strong argument that a foundation for such a claim exists.

If it were permitted, how would a products liability claim against a chatbot work? Product liability claims may be rooted in negligence, strict liability, or breach of warranty.[83] However, the general defamation principle that plaintiffs must prove at least negligence would suggest that for reputational harms caused by chatbots, plaintiffs would be required to bring the claim under negligence as opposed to strict liability.

Although claims arising under products liability law differ (sometimes widely) by jurisdiction,[84] there are three main theories of liability under products liability law: design defect, manufacturing defect, and failure to warn/instruct. Depending on the circumstances of an individual case, when a chatbot has produced false

---

[82] *See* Nw. Arkansas Masonry v. Summit Specialty Prod., 31 P.3d 982, 987 (Kan. 2001) (noting that damages arising as a "result of the failure of the product to perform to the level expected by the buyer . . . is the core concern of traditional contract law").

[83] 63 AM. JUR. 2D PRODUCTS LIABILITY § 5 (2012); David G. Owen, *Manufacturing Defects*, 53 S.C. L. REV. 851, 860 (2002) (noting that most states allow plaintiffs to pursue recovery under negligent manufacturing claims).

[84] LOUIS FRUMER ET AL., 1 PRODUCTS LIABILITY, § 8.01 (2023) (quoting Hon. Justice Neely of the West Virginia Supreme Court calling products liability law "the peculiarly American System of fifty, uncoordinated, separate schemes of tort law coexisting within one industrial nation").

speech that harms someone's reputation a claim could be plausible under each of those theories. I will look at each in turn.

B. Designing Around Defamation

When a product has an inherent defect in its design that makes it dangerous to consumers, a plaintiff may be able to claim a design defect.[85] Having a design defect doesn't mean the product doesn't serve its purpose—a product might work well in most situations but still be unreasonably dangerous to use.[86] Consider a toaster oven that was designed to broil food on high heat, but the high heat actually caused components of the toaster oven to catch on fire. This would be a defective design. Design defects arise from a variety of factors, including inadequate testing, lack of safety features, or failure to reasonably account for foreseeable risks.

Typically, plaintiffs can establish design defect only when they can prove that there is another (even hypothetical) alternative design that would be safer than the original, but as economically feasible and practical.[87] But as long as the design is reasonably safe, the defendant is not obligated to design the safest possible product, or even one as safer than it has designed.[88]

In the context of speech harms caused by chatbots, a design defect could exist if the model was designed in a way that made it likely to generate defamatory statements. This would require an examination into how the model was programmed and how it operates. Its training data, algorithms, product testing, and programmer decisions would all be relevant to a determination of whether there was a defective design. Although there are certainly others, I'll offer a few examples where a plaintiff would have an argument that the design was defective:

---

[85] Barker v. Lull Eng'g Co., 573 P.2d 443, 454 (Cal. 1978) ("[A] product may be found defective in design if the plaintiff demonstrates that the product failed to perform as safely as an ordinary consumer would expect when used in an intended or reasonably foreseeable manner.").

[86] *Id.* ("[A] product may be found defective in design, even if it satisfies ordinary consumer expectations.").

[87] *See* RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 2(b) (1997) ("The foreseeable risks of harm posed by the product could have been reduced or avoided, by the adoption of a reasonable alternative design by the seller or other distributor, or a predecessor in the commercial chain of distribution, and the omission of the alternative design renders the product not reasonably safe.").

[88] Cornstubble v. Ford Motor Co., 532 N.E.2d 884 (Ill. App. Ct. 1988); Hagans v. Oliver Machinery Co., 576 F.2d 97 (5th Cir. 1978) (applying Texas law).

- Programmers selected a flawed dataset, perhaps by pointing the algorithm at a dataset rife with false content.
- Programmers prioritized generating sensational or controversial content over the accuracy of such content.
- Programmers did not take steps to reduce the likelihood of hallucinations—for example, if developers designed a chatbot to scrape the internet for information and generate articles based on that information without using classifiers on known risk vectors to reduce the tendency of the models to hallucinate.
- Programmers did not test the chatbot to see how it responded to prompts asking it for false or controversial content.
- The design lacks adequate controls to prevent defamatory speech from being generated. For example, if the chatbot was programmed to generate speech based on user input/prompts, it is foreseeable that some user prompts would "lead" the chatbot to produce statements that could cause damage to someone's reputation. If the developers know that the algorithm is quite prone to hallucination and elected not to put controls in place (e.g., programming the chatbot to refuse to answer certain prompts, or supplying disclaimer language to respond to particular types of prompts), the design could be defective.[89]

With all of the above examples, the court would have to consider whether there were feasible alternative designs that could have prevented the defamatory speech from being generated. For example, if developers could have created a dataset that didn't contain false content, designed the chatbot to verify the accuracy of the information it scraped, designed it to remove potentially defamatory statements, or designed it with more robust controls to prevent defamatory speech from being generated, those would constitute feasible design alternatives.

Much of design defect law hinges on the concept of foreseeability. An important consideration in all products liability claims against chatbot developers,

---

[89] *See* Derek E. Bambauer & Mihai Surdeanu, *Authorbots*, 3 J. FREE SPEECH L. 375, 379 (2023) (noting that "[n]ormatively undesirable outputs, such as false or defamatory ones, are more readily generated by inputs that push ChatGPT in that direction—in other words, by leading questions. Defamatory blossoms often result from defamatory seeds").

including design defects, would be what it means to be a reasonable consumer. Reasonable consumers use products in foreseeable and expected ways, and do not engage in negligent or reckless behavior that contribute to the injury. When consumers do not behave reasonably and that behavior was not foreseeable, it often cuts off liability for manufacturers and sellers.[90]

What does it mean to be a reasonable consumer in this circumstance? Would it be reasonable for a consumer to use the chatbot to look up information about people and rely on the information provided? Would a reasonable consumer use a chatbot to help them learn new skills, improve their knowledge on certain topics, or make informed decisions? Would a reasonable consumer understand enough about how the product works to know that reliance on its output could lead to liability? In this scenario, reasonable consumers will likely have a very limited (if any) understanding of the technology, and may not be aware that the chatbot might produce false information—or that repeating that false information might lead to liability.

In any case, even where the consumer may be acting unreasonably, there are clearly foreseeable harms associated with chatbots, some of which have already been acknowledged by developers.[91] It is foreseeable that users will request illicit advice, for example, or that chatbots will produce false information. It is also foreseeable that there will be problems with the data, given that the datasets are so large. (Consider that newer versions of ChatGPT can scrape from the web to respond to user prompts.) Because of this, developers have a responsibility to build in appropriate safety measures so that problems in the data do not lead to harmful speech by the chatbot.

An additional wrinkle emerges when one considers that chatbots are products designed to respond to unique user prompts, so the developer is designing a product that will respond to and interact with users in ways that developer could not have contemplated. So while some "dangerous" uses may be foreseeable and can be mitigated ahead of launch, others may emerge as users interact with the product.

---

[90] Indian Brand Farms v. Novartis Crop Prot., 617 F.3d 207, 225 (3d Cir. 2010) ("[A] manufacturer is not liable for damages where a person misuses the product, unless that misuse was objectively foreseeable. . . . [W]here the use of the product is beyond its intended or reasonably anticipated scope, an injury resulting from that use is not . . . probative of whether the product was fit, suitable, and safe." (cleaned up)).

[91] *See* Wiggers, *supra* note 30.

OpenAI anticipated some of this and put safeguards in place to guard against harms. For example, the developers anticipated that users might ask ChatGPT to "get dirt" on a particular person, and programmed it to refuse to respond to such requests by explaining that it cannot verify criminal history and as such it would be inappropriate to spread false information or defame someone's character.[92] What the developers *didn't* fully anticipate is that users would find workarounds. "[A]ll you have to do is ask for that defamation in a language other than English, et voilà: coherent articles about notorious villains, and their entirely made-up criminal histories—which it'll happily translate back into English, should you ask it to."[93]

Even if a plaintiff can prove that a design is defective, courts will sometimes employ a risk/utility test to see if the product's utility outweighs its inherent risk of harm.[94] Under this test, courts "balance the risks of the product as designed against the costs of making the product safer."[95] Thus, courts would consider the costs of a particular precaution and evaluate whether they were less than the precaution's safety benefits.

For example, it might be possible for developers to improve the quality of their dataset by individual review of each piece of data, reducing the likelihood that the chatbot produces defamatory or otherwise harmful speech. But doing so would require significant time and cost. Any increase in accuracy of the chatbot's speech gained by this effort might not outweigh the time and expense of implementation, especially given that costs of making the product safer include any loss of product

---

[92] Harrison, *supra* note 5; *see also* Kevin Roose, *The Brilliance and Weirdness of ChatGPT*, N.Y. TIMES (Dec. 5, 2022) (explaining how users have found ways to circumvent the guardrails put in place by ChatGPT).

[93] *Id.*

[94] *Indian Brand Farms,* 617 F.3d at 225 ("The decision whether a product is defective because it is not reasonably fit, suitable and safe for its intended purposes reflects a policy judgment under a risk-utility analysis that seeks to determine whether a particular product creates a risk of harm that outweighs its usefulness."); David G. Owen, *Design Defects,* 73 MO. L. REV. 291, 307 (2008) ("The risk-utility test is the principal standard for judging the safety or defectiveness of a product's design.").

[95] DAN B. DOBBS ET AL., THE LAW OF TORTS § 456 [The Risk-Utility Test for Design Defects] (2d ed.).

utility.[96] This might be a circumstance where courts would hold that the risk is justifiable given the utility of chatbots, and hold that no design defect exists.

On the other hand, courts might determine that the risk/utility analysis demands designers employ controls to prevent defamatory speech from being generated in the first place. Programming the chatbot to refuse to answer certain "loaded" questions—or to answer questions carefully, with programmed qualifiers—is of relatively low cost to the developer, but greatly reduces the likelihood of harm to third parties.

Some risk mitigation efforts, though expensive, may still be necessary under such an inquiry. Involving humans in the "fine-tuning" phase is a significant cost, but it reduces hallucinations and renders the product better able to respond to certain prompts and refuse others (such as direct requests for illicit advice) altogether. The court might also find the magnified risk of harm relevant. These models are capable of producing content at a rapid pace and on a massive scale. This means that a single design defect in an AI system can lead to a large number of defamatory statements being generated and disseminated because of the scale at which the chatbots operate.

Ultimately, even if plaintiffs can prove that the design of the chatbot caused them harm, liability would not be appropriate if the product's utility outweighs its inherent risk of harm.

C. Manufacturing Defamation

Manufacturing defects occur when an individual product departs from its intended design because of a flaw in the manufacturing process.[97] Simply put, there is a gap between what the manufacturer intended to make and what it actually produced, and as such the product does not conform to the manufacturer's own specifications or requirements.[98] When it comes to generative AI, the algorithm *is* the

---

[96] *Id.*

[97] Frye v. Biro Mfg. Co., No. C10-0192-JCC, 2011 WL 6013775, at *5–6 (W.D. Wash. Dec. 2, 2011) ("The case law for manufacturing defect recognizes that a manufacturer is liable if the defect results from a failure in the manufacturing process to deliver the product as intended."); *see also* RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 2(a) (1997).

[98] RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 2(a) (1997) ("[A] manufacturing defect [exists] when the product departs from its intended design even though all possible care was exercised in the preparation and marketing of the product.").

product. Thus, the developer doesn't sell or license individual chatbots, but rather licenses or makes available the one chatbot to all users.

For this reason, claims based on this theory may be less likely to exist (and succeed) than design defects, where there is a clear application of the law. It is possible that a plaintiff could make a claim for manufacturing defect on the basis that a logical error in the algorithm caused it to generate defamatory statements. The argument would be that the defect was a manufacturing defect because it was a problem with the *actual product*—the algorithm—rather than a problem with the product's design. Such an interpretation would make it difficult to untangle manufacturing defects from design defects and would therefore likely be unsuccessful.

An interesting query arises with generative AI and manufacturing defects, however, because the code programmed by the developers often evolves in ways the programmers could not have predicted.[99] Unlike most products, generative AI models are not finished products when they launch. One of their key attributes, in fact, is their ability to "evolve," and it has the capability to change far beyond what its original programmers anticipated.[100]

If defects introduced into the product at the replication and distribution phase would be deemed manufacturing defects, would the evolution of the code into a new product constitute a manufacturing defect for which the developer should bear liability? There is a compelling argument that the developer should bear that liability because it is the least-cost avoider.[101] Where the AI evolves in such a way that it

---

[99] Edd Gent, *Artificial Intelligence Is Evolving All By Itself*, SCIENCE (Apr. 13, 2020), https://www.science.org/content/article/artificial-intelligence-evolving-all-itself; Stephen Ornes, *The Unpredictable Abilities Emerging From Large AI Models*, QUANTA MAGAZINE (Mar. 16, 2023), https://perma.cc/TPW2-M3RF.

[100] Sarah Brown, *Machine Learning, Explained*, MIT SLOAN SCHOOL OF MANAGEMENT (Apr. 21, 2021), https://perma.cc/7WVT-H7QN ("Machine learning takes the approach of letting computers learn to program themselves through experience."); *The Impact of Artificial Intelligence on the future of Workforces in the European Union and the United States of America*, WHITE HOUSE (Dec. 5, 2022), https://perma.cc/L52N-QV6K ("The power of AI comes from its use of machine learning, a branch of computational statistics that focuses on designing algorithms that can automatically and iteratively build analytical models from new data without explicitly programming the solution.").

[101] This may be a scenario where they are *not* the least-cost avoider if we acknowledge that there is an overall net-positive from the use of this technology. If that is the case, is the reality that there may be no appropriate defendant to bear the cost of this harm and that it stays with the plaintiff? Or

renders the product more useful or effective at rendering results, the developer will be able to reap those rewards. The converse should be true. If the evolution renders the product harmful, the company should bear that liability as well. It is unclear how courts would resolve this question.

D. Failure to Warn Against Defamation

A failure to adequately warn or instruct users about the inherent dangers associated with using chatbots may give harmed plaintiffs another opportunity for recourse. An assumption in products liability law is that manufacturers and sellers are in a better position than users to anticipate inherent dangers associated with their products and are thus in a better position to warn of these dangers. For this reason, manufacturers and sellers have a duty to warn when they are aware (or should be) that their product is dangerous, the danger is present when the product is used in the usual and expected manner, and the danger is not obvious or well known to the user.[102] Third parties injured by products may allege failure to warn claims against manufacturers for failing to warn the purchasers about possible dangers.[103] When manufacturers/sellers fail to provide *any* warnings, fail to provide *adequate* warnings, or fail to adequately instruct users on how to use the products and avoid harm, they can face liability for failure to warn/instruct.[104] Warnings

---

should companies bear the legal responsibility for products they create even when those products evolve in ways that were not foreseeable.

[102] *See* Billiar v. Minnesota Mining & Mfg. Co., 623 F.2d 240, 243 (2d Cir. 1980).

[103] *See, e.g.*, Erie Ins. Co. v. W.M. Barr & Co., 523 F. Supp. 3d 1, 13 (D.D.C. 2021) (failure to warn claim permitted against chemical manufacturers whose products allegedly caused an explosion in insured's property); Isatou Bah v. Nordson Corp., No. 00CIV9060DAB, 2005 WL 1813023, at *15 (S.D.N.Y. Aug. 1, 2005) (allowing negligence claim to be brought against manufacturer by an employee harmed by another's use of glue dispensing machine, noting "New York Courts have allowed a failure to warn claim by a bystander plaintiff injured by a product produced by a defendant manufacturer but used by third parties"); La Paglia v. Sears Roebuck & Co., 531 N.Y.S.2d 623, 625–26 (2d Dep't 1988) (bystander plaintiff allowed to bring failure to warn claim against manufacturer of lawnmower owned and driven by co-defendant that caused plaintiff's injury); Mack v. Ford Motor Co., 669 N.E.2d 608, 611 (Ill. App. Ct. 1996) (allowing failure to warn claim against manufacturer brought by the family of third-party decedent).

[104] PLIVA, Inc. v. Mensing, 564 U.S. 604, 611 (2011) ("a manufacturer's duty to warn includes a duty to provide adequate instructions for safe use of a product"); *see also* RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 2(c) (1997) ("[I]s defective because of inadequate instructions or warnings when the foreseeable risks of harm posed by the product could have been reduced or avoided by the provision of reasonable instructions or warnings by the seller or other distributor.").

must also be conspicuous—in a place where users are going to be able to readily find and read them.[105]

Does a duty to warn arise in the context of chatbots? Given the way generative AI models function, it is entirely foreseeable that a dataset which scrapes from the open web would contain false or harmful information. Because the algorithm recognizes, summarizes, and predicts text based on the dataset, it is clear that where the dataset contains inaccurate information, so too might the algorithm's output. It is also entirely foreseeable that an algorithm that is trained to recognize, summarize, translate, predict, and generate text—but that does not understand the meaning of what it produces—will sometimes predict words in a sequence that results in a false and potentially defamatory statement even when the dataset is entirely accurate. Indeed, programmers have acknowledged these concerns. The developers at OpenAI have acknowledged that the ChatGPT product is still in its nascent stages and a work in progress. "It's a mistake to be relying on it for anything important right now," OpenAI Chief Executive Sam Altman tweeted.[106] "We have lots of work to do on robustness and truthfulness."[107] Despite programmers' efforts to reduce harmful speech, the likelihood still exists that chatbots may produce defamatory and other harmful statements. Thus, a warning would be required to alert users to this possibility.

Once a duty to warn arises, it is not enough for the manufacturer to provide *any* warning. Instead, it must be an *adequate* warning.[108] What constitutes an

---

[105] Town of Bridport v. Sterling Clark Lurton Corp., 693 A.2d 701, 704 (Vt. 1997) (explaining that to be adequate, a warning about a product must be displayed so as to catch the eye of a reasonably prudent person; bold and prominent warnings of the dangers of fire and spontaneous combustion were sufficient to warn of the danger of fire that resulted when materials soaked with linseed oil and gum turps spontaneously combusted; there was no showing that the warnings were not sufficiently conspicuous); Maneely v. Gen. Motors Corp., 108 F.3d 1176, 1179 (9th Cir. 1997) ("When a manufacturer is or should have been aware that a product is unreasonably dangerous absent a warning and such warning is feasible, the manufacturer will be held strictly liable if it fails to give an appropriate and conspicuous warning."); McEwen v. Ortho Pharm. Corp., 528 P.2d 522, 530 (Ore. 1974); RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 2 cmt. j (1998).

[106] Sam Altman (@sama), TWITTER (Dec. 10, 2022, 7:11 PM), https://twitter.com/sama/status/1601731295792414720.

[107] *Id.*

[108] AM. L. PROD. LIAB. 3D § 33:1 (1997) ("Providing an inadequate warning is no better than providing no warning at all.").

adequate warning or an adequate instruction can be a highly subjective and fact-intensive evaluation. However, courts typically require that the warning should attract the user's attention, explain the dangers of the product and how to safely use it, taking into account the knowledge and expertise of those who may be reasonably expected to use the product.[109] Manufacturers and sellers may demonstrate adequacy of the warning by showing that they "clearly alerted the user to avoid certain uses of the product that appear to be normal and reasonable or that the warning informed the user of the particular injury alleged."[110] When industries have standards for providing warnings and guidelines, following or exceeding such guidelines is often evidence of adequacy. This is unlikely to be the case in a field such as generative AI, which it is rapidly evolving and in which no industry guidelines exist.

For LLMs such as chatbots, adequacy of warning might mean an alert that clearly warns users to the potential harms of using the chatbot—that it can produce false information, that it can produce information that if reproduced elsewhere could be the basis for legal liability. Perhaps instead of posting a passive warning that users can easily miss or ignore, developers could require users to "accept" the risks by restricting access to the chatbot until users click a box indicating that they agree and understand certain enumerated potential harms associated with the product.

To succeed on a failure to warn/instruct claim, plaintiffs must also prove that the failure to warn, or the failure to adequately warn, was the proximate cause of

---

[109] *See, e.g.*, Glorvigen v. Cirrus Design Corp., 796 N.W.2d 541, 550 (Minn. Ct. App. 2011), *aff'd*, 816 N.W.2d 572 (Minn. 2012); Kendall v. Hoffman-La Roche, Inc., 36 A.3d 541, 554 (N.J. 2012) ("An adequate product warning or instruction is one that a reasonably prudent person in the same or similar circumstances would have provided with respect to the danger and that communicates adequate information on the dangers and safe use of the product, taking into account the characteristics of, and the ordinary knowledge common to, the persons by whom the product is intended to be used . . . ."); Daimlerchrysler Corp. v. Hillhouse, 161 S.W.3d 541, 548–49 (Tex. App. 2004) ("An adequate warning is one given in such form that (1) it could reasonably be expected to catch the attention of the reasonably prudent person in the circumstances of its use; and (2) its content is comprehensible to the product's average user and conveys a fair indication of the nature and extent of the danger, if any, and how to avoid it.") *review granted, judgment vacated, and remanded by agreement*, No. 05-0289, 2006 WL 8473582 (Tex. Jan. 13, 2006).

[110] 4D N.Y. PRAC., COM. LITIG. IN NEW YORK STATE COURTS § 107:41 [Warning Was Adequate] (5th ed. 2022).

their injury.[111] In the case of a chatbot that produced defamatory text, application of this step would likely require the injured party to prove that the chatbot user would not have used the chatbot (or would have used it differently) if an adequate warning had been given. [112] Let's return to our example of the hiring manager who asks a chatbot about a candidate for a position, and the chatbot produces a response falsely stating that the candidate has a history of fraud and embezzlement. Does a warning that results are not reliable—and may in fact be false—avoid the harm? If the plaintiff can prove that the hiring manager would have heeded the warning by not using the chatbot, or at least not taking information it provided at face value, this could support plaintiff's case.

In the case of ChatGPT, OpenAI offers multiple warnings to users. On the landing page, three product limitations are noted in the center of the page, directly above the field for users to enter a prompt. These include warnings that the product "May occasionally generate incorrect information" and "May occasionally produce harmful instructions or biased content."[113] Directly below the prompt field a second warning reads that "ChatGPT may produce inaccurate information about people, places, or facts."[114] Though the warnings make clear that the chatbot may produce inaccurate information, they stop short of alerting users that the chatbot can generate information that if republished elsewhere could be the basis for legal liability. Without this clarity, are the warnings it offers adequate?

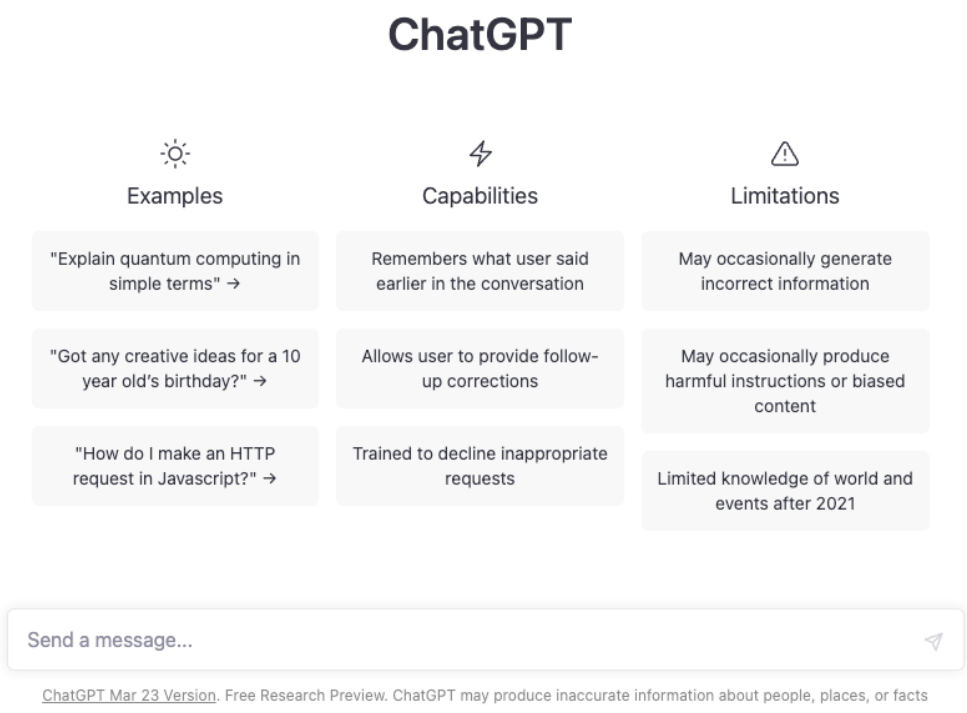
---

[111] Eck v. Parke, Davis & Co., 256 F.3d 1013, 1017 (10th Cir. 2001) ("To recover in a failure to warn case, a plaintiff must establish both cause-in-fact (that the product in question caused the injury) and proximate cause (that the manufacturer of the product "breached a duty to warn of possible detrimental reactions"). To qualify as a proximate cause of the injury, the breach of a duty or failure to warn must be a substantial contributing factor in bringing about the harm in question.").

[112] *See* Volokh, *supra* note 46, at 500 (noting that "[e]ven if the AIs' *users* are seen as waiving their rights to sue based on erroneous information when they expressly or implicitly acknowledge the disclaimers, that can't waive the rights of the *third parties* who might be libeled").

[113] *ChatGPT*, OpenAI, https://chat.openai.com/chat.

[114] *Id.*



OpenAI offers further warnings in its terms of use, where it explains that

> Artificial intelligence and machine learning are rapidly evolving fields of study. We are constantly working to improve our Services to make them more accurate, reliable, safe and beneficial. Given the probabilistic nature of machine learning, use of our Services may in some situations result in incorrect Output that does not accurately reflect real people, places, or facts. You should evaluate the accuracy of any Output as appropriate for your use case, including by using human review of the Output.[115]

But are the warnings offered in its terms of use sufficient? Warnings offered in terms of use may be considered as part of the overall warning package, but they may not be sufficient on their own to discharge the duty to warn. The adequacy of a warning will depend on the nature and extent of the potential risks associated with the product, as well as the characteristics of the intended user population.[116]

It bears mention that warnings alone (even if required) would not be adequate to guard against a claim of defamation. As Eugene Volokh notes, "[d]efamation law has long treated false, potentially reputation-damaging assertions about people as actionable even when there's clearly some possibility that the assertions might be

---

[115] *Terms of Use*, OpenAI (Mar. 14, 2023), https://openai.com/policies/terms-of-use.

[116] Wooderson v. Ortho Pharm. Corp., 681 P.2d 1038, 1055 (Kan. 1984).

false."[117] Even when accompanied by clear warnings that the information is possibly false, that is unlikely to be a bar to a defamation claim. [118]

Because of the way chatbots operate, it seems evident that warnings or instructions are necessary to prevent foreseeable harms. Depending on the warnings given—or not—by a developer, a plaintiff could have a claim grounded in failure to warn.

IV. Evaluating a Defamation Claim Through a Products Liability Lens

If courts reject claims for harm to reputation pled under products liability law, plaintiffs may have no choice but to bring them under defamation law. As discussed in Section II, this creates a challenge for evaluating the element of fault. This element would require plaintiffs—depending on their status as a private citizen, public official, or public figure—to prove that the defendant acted with either negligence or actual malice. In cases where plaintiffs must prove negligence, they are typically required to show that the defendant did not use reasonable care in ascertaining the truth or falsity of the statements.[119] Where they must prove actual malice, they are required to show that the defendant made the statement with knowledge it was false, or with reckless disregard as to the truth or falsity of the statement.[120]

Proving that the developer had *any* particular level of care with respect to the truth or falsity of specific statements made by the chatbot will be a challenge because the developer designed the chatbot to respond to user prompts autonomously. Its role in the preparation of the publication was to program and train the chatbot, not to draft specific statements. Given this, the more appropriate inquiry would be to look at the mental state of individuals within the organization responsible for programming and training, even though as discussed above they have no

---

[117] Volokh, *supra* note 46, at 500.

[118] *Id.*

[119] Pacitti v. Durr, 310 F. App'x 526, 528 (3d Cir. 2009) ("[P]laintiffs needed to prove that the defendant published defamatory material in a negligent manner . . . negligence in this context is the publication of information with a want of reasonable care to ascertain the truth."); Straw v. Chase Revel, 813 F.2d 356, 359 (11th Cir. 1987) (holding that negligent conduct is "a failure to exercise that degree of care exercised under the same or similar circumstances by ordinarily prudent persons"); Kendrick v. Fox Television, 659 A.2d 814, 823 (D.C. 1995); Gannett Co. v. Kanaga, 750 A.2d 1174, 1181 (Del. 2000).

[120] New York Times Co. v. Sullivan, 376 U.S. 254, 280 (1964).

role in the preparation of the publication. This is where products liability law provides a useful framework, particularly in the evaluation of negligence.

Because products liability claims may be brought under a theory of negligence, there is an entire body of law informing the analysis of whether a manufacturer used reasonable care in programming, training, and deploying the chatbot. If a plaintiff could prove that the defendant was negligent in its design, manufacture, or warnings, that should suffice to prove the element of fault in a defamation action. If there were flaws in the product's design, for example, because the programmers prioritized generating sensational or controversial content over accurate and non-biased content, that might suggest the developer failed to use reasonable care in designing a chatbot that could prepare its own speech. Or perhaps plaintiffs could prove that the developers did not include adequate warnings because they failed to alert users that the chatbot could produce false information, which could likewise demonstrate carelessness.

In either circumstance, if plaintiffs could demonstrate that the defendant was negligent with the standards established in products liability law, this should be enough to prove negligence for the purposes of defamation law. The developers may not have prepared the publication, but they enabled the chatbot *to* publish, so the body of law built to analyze when designers and manufacturers bear liability for their products supplies a helpful analogy.

Where plaintiff is required to prove actual malice, products liability law offers less of a plug-and-play analogy but may still provide useful guidance. Plaintiffs can prove actual malice either by showing that the defendant knew the published statement was false or that the defendant had a reckless disregard for whether it was true or false.[121]

When could developers have knowledge that their chatbots are publishing false and defamatory statements? Few circumstances are likely to emerge. As a practical matter, there is an incentive for developers to produce a chatbot that produces reliably truthful responses because that is better for business. It's highly unlikely that developers would program the chatbot to respond to user prompts with false and defamatory statements, but that of course would constitute knowledge of falsity. So too would a situation where developers are aware that the chatbot consistently produces false information in response to certain types of prompts. For example, if the

[121] *Id.* at 280; McDougal v. Fox News Network, LLC, 489 F. Supp. 3d 174, 185 (S.D.N.Y. 2020).

programmers know that the chatbot unfailingly produces criminal records for people who have no criminal record when given certain prompts, the developers have constructive knowledge that the chatbot is producing false (and likely defamatory) speech.[122]

But what if the chatbot doesn't *consistently* produce criminal records for people who have none, and only does it occasionally? In defamation cases, a defendant would be found reckless if it knew that there was a substantial risk that the statement was false, but made it anyway.[123] It is arguable that if the programmers know that this happens—either because their own testing has revealed the issue or others have pointed it out—that would meet the threshold of recklessness.

Once aware of the risk, a publisher must take steps to verify the accuracy of the statements. By analogy to products liability law, recklessness could exist if the developer is aware that there is a substantial risk that the chatbot is publishing false and defamatory statements but does nothing to mitigate this risk.[124] Consider an example where a trainer knowingly inserts false information into the training corpus. There is no guarantee that the chatbot will produce false speech consistent with that training data, but there is a serious risk that it would. Or if the developer is told by a user that the chatbot is producing false statements about an individual and it ignores the notice despite the warning about the chatbot's speech, this would be "textbook" recklessness.[125]

Chatbot developers will be generally aware that there is a risk their products will produce false and defamatory speech. After all, they have trained an algorithm to predict answers to user prompts based on a massive (and at least partly unverified) dataset. But it wouldn't be right to characterize this awareness as reckless disregard so long as the developers take steps to mitigate this risk. The decisions developers make in design, manufacturing, and warnings/instructions are important

---

[122] Schafer v. Time, Inc., 142 F.3d 1361, 1366 (11th Cir. 1998) (holding that "'actual malice' refer[s] to the speaker's actual or constructive knowledge regarding the truth of the statement").

[123] *Id.*

[124] Volokh, *supra* note 46, at 518 (discussing the possibility of a notice and blocking model and noting that AI companies "should be capable of doing something to diminish the repetition of libelous allegations to which they have been specifically alerted").

[125] *See generally id.* at 514–21.

not only because they impact the safety of the product, but also as indicators of how seriously developers worked to mitigate known risks.

The framework of products liability law is clearly a more apt analogy for private plaintiffs who will be required only to prove negligence but may offer courts some insights as to recklessness as well.

Conclusion

The emergence of chatbots and their potential for speech harm poses a challenge for applying traditional defamation law. While the default assumption is to pursue a case for defamation, assigning fault to the chatbot or its developers is a complicated task. Products liability law, however, presents a viable alternative theory of liability—or at least a framework—to address the harms caused by chatbot-generated speech. This area of law is well-suited to adapt to emerging technologies like generative AI, and requires that those in the best position to eliminate risks and harms should bear the legal responsibility for injury caused by their products. Ultimately, a products liability framework for assessing fault would compensate those injured by chatbots, deter the launch of chatbots that are unsafe, and financially sanction those manufacturers who offer such unsafe chatbots to the public.