



INEVITABLE ERRORS: DEFAMATION BY HALLUCINATION IN AI REASONING MODELS

Lyrisa Lidsky * & *Andrew Daves* **

Introduction.....	478
I. <i>Walters</i> and Foundation Model Defamation by Hallucination	489
A. “I Apologize for the Confusion”: The <i>Walters</i> Decision	489
1. “Actual facts”	490
2. Scierter—negligence and actual malice.....	495
3. Presumed damages	503
4. The path not taken in <i>Walters</i> : Is ChatGPT a publisher or a distributor?.....	505
II. A Moving Target: OpenAI’s Evolution from Foundation Models to Reasoning	517
A. A Chain-of-Thought Timeline.....	519
B. Hallucinations as Inevitable (and Sometimes Valuable!) Errors.....	531
III. Applying Defamation Law to LLM Reasoning Models: Evaluating Policy Tradeoffs and Existing Proposals	538

* Lyrisa Lidsky is Raymond & Miriam Ehrlich Chair in U.S. Constitutional Law, University of Florida Fredric G. Levin College of Law. This paper benefited from comments contributed by students, faculty, and fellows who attended the Yale Information Society Project workshop on February 18, 2025; faculty at the University of Minnesota School of Law on February 24, 2025; participants at the Yale Freedom of Expression Scholars Conference in April 2025, and participants at the Tübingen Conference on AI and Law. We thank Tom Cotter, Lili Levi, Federico Jorge Gaxiola Lappe, James Grimmerman, Jane Bambauer, Chris Hampson, Elisabeth (Liz) Lear, Berta Hernandez-Truyol, Jiaying Jiang, RonNell Andersen Jones, Helen Norton, Hari Osofsky, Blake Reid, Margot Kaminski, Douglas Spencer, Scott Skinner-Thompson, Christina Koningisor, Benjamin Duren Lidsky, and Grace Chisholm for useful conversations and comments that improved the quality of this draft.

** Andrew Daves is a third-year law student at the University of Florida Levin College of Law.

A.	Six Insights from the Changing LLM Landscape.....	538
B.	Communications Policy and the Defamation-by-Hallucination Problem.....	540
1.	Compensate defamation victims for provable harms.....	541
2.	Correct the record and prevent future errors	543
3.	Incentivize model transparency to reduce future harms.....	544
4.	Incentivize innovations for safety and accuracy	546
5.	Incentivize the exercise of reasonable care by AI producers	547
6.	Center liability on the cheapest cost avoider	548
IV.	Applying Liability Rules to Reasoning Model Defamation	549
A.	The Limits of Treating Hallucinations as Negligent-Design Defects	550
B.	The Limits of Notice-and-Blocking Based Liability	554
C.	The Case for Imposing Statutory Duties on AI Producers	558
D.	LLM Producers as Quasi-Distributors—and the Special Instance of Chains of Thought Hallucinations.....	560
E.	Dangers of Imposing Strict Liability.....	564
	Conclusion	567

INTRODUCTION

Artificial general intelligence is “probably the greatest threat to the continued existence of humanity.”¹ Or so claims OpenAI’s Chief Executive Officer Sam Altman.² In a seeming paradox, OpenAI defines its mission as ensuring “that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity.”³

¹ John Herrman, *What Ever Happened to the AI Apocalypse?*, INTELLIGENCER (June 4, 2024), <https://perma.cc/2LR7-RKJF> (quoting Sam Altman).

² *Id.*

³ *About*, OPENAI, <https://perma.cc/W4NS-YNPF>. Artificial general intelligence (AGI) represents a system that can “efficiently acquire *new* skills outside of its training data.” *ARC-AGI*, ARC PRIZE, <https://perma.cc/V56Q-ZYT9> (emphasis in original). AGI is an important benchmark for

Whether artificial general intelligence becomes a universal boon⁴ or an existential threat⁵—or both—there is general agreement concerning its revolutionary potential. Indeed, Microsoft founder Bill Gates has called it “as fundamental an advance as the personal computer or the internet,”⁶ and Google CEO Sundar Pichai has predicted that it will “bring about a more profound shift in human life than electricity or Promethean fire.”⁷

Thus far, AI systems are not *generally* smarter than humans. Not yet.⁸ Large Language Models (LLMs), however, are advancing at a startling pace. LLMs use artificial intelligence to synthesize massive amounts of textual data and then predict text and generate responses to users in “natural” human language.⁹ On a scale measuring the progress of LLMs toward general intelligence,¹⁰ OpenAI’s flagship

researchers because it is “typically intertwined with a notion of ‘emergent’ properties, i.e. capabilities not explicitly anticipated by the developer. Such capabilities offer promise, perhaps including abilities that are complementary to typical human skills, enabling new types of interaction or novel industries.” Meredith Ringel Morris et al., *Position: Levels of AGI for Operationalizing Progress on the Path to AGI* (Sep. 24, 2025), <https://perma.cc/AS25-GCKL>.

⁴ According to technologist Vinod Khosla, “AI promises a future of unparalleled abundance”—but only if we regulate it to guarantee that “democratic values prevail.” Vinod Khosla, *A Roadmap to AI Utopia*, TIME (Nov. 11, 2024), <https://perma.cc/P2JL-76R4>.

⁵ An example of the apocalyptic threat is that AI will make us all dumber. See, e.g., Nataliya Kosmyna et al., *Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task* (June 10, 2025), <https://perma.cc/A7MM-QX8S>.

⁶ Ross Andersen, *Does Sam Altman Know What He’s Creating?*, ATLANTIC (July 24, 2023), <https://perma.cc/R3GH-FEUL>.

⁷ *Id.*

⁸ Morris et al., *supra* note 3, at 1 (“Some experts believe that ‘sparks’ of AGI are already present in the latest generation of large language models (LLMs); some predict AI will broadly outperform humans within about a decade; some even assert that current LLMs are AGIs.” (emphasis in original) (internal citations omitted)).

⁹ Cole Stryker, *What Are Large Language Models (LLMs)?*, IBM (Sep. 10, 2025), <https://perma.cc/V86M-MYHD>. See generally Timothy B. Lee & Sean Trott, *A Jargon-Free Explanation of How AI Large Language Models Work*, ARS TECHNICA (July 31, 2023), <https://perma.cc/8UN5-9E9K>.

¹⁰ The ARC-AGI test has measured LLM progress toward general intelligence in AI models since 2019. ARC PRIZE, *supra* note 3.

model, GPT-3 scored zero percent in 2020.¹¹ Just five years later, OpenAI's o3-preview scored between 75% and 88%.¹² Meanwhile, OpenAI competitors such as Anthropic, Google, and DeepSeek are likewise racing to deliver on the promise of "systems that can think and act rationally in ways that mirror human behavior and intelligence."¹³

Even as LLM models make progress toward general intelligence, there are already AI systems that have exceeded human performance on narrow, clearly scoped tasks.¹⁴ For example, chess engines have been performing at superhuman levels for years, and AI models can now help detect breast cancer far earlier than human experts—and the models continue to improve.¹⁵ Meanwhile, OpenAI's o1 reasoning model has an LSAT score higher than the median student admitted to the law schools at Harvard, Yale, and Stanford.¹⁶

¹¹ Katharina Buchholz, *The Extreme Cost of Training AI Models*, FORBES (Aug. 26, 2024), <https://perma.cc/CF6G-KNYJ>.

¹² Brian Buntz, *Just How Big of a Deal Is OpenAI's o3 Model Anyway?*, R&D WORLD (Dec. 23, 2024), <https://perma.cc/NWM8-DUU9>.

¹³ See Jared Schroeder, *Saving the Marketplace from Market Failure: Reorienting Marketplace Theory in the Era of AI Communicators*, 28 WM. & MARY BILL RTS. J. 689, 696 (2020) (observing that "scholars have defined AI as systems that can think and act rationally in ways that mirror human behavior and intelligence").

¹⁴ François Chollet, *On the Measure of Intelligence* (Nov. 25, 2019), <https://perma.cc/PGQ4-7K77>. See also Morris et al., *supra* note 3, at 5 (chart).

¹⁵ Veronica Hernstrom et al., *Screening Performance and Characteristics of Breast Cancer Detected in the Mammography Screening with Artificial Intelligence Trial (MASAI): A Randomised, Controlled, Parallel-Group, Non-Inferiority, Single-Blinded, Screening Accuracy Study*, 7 LANCET DIGIT. HEALTH e175 (2025); Nora Eisemann et al., *Nationwide Real-World Implementation of AI for Cancer Detection in Population-Based Mammography Screening*, 31 NATURE MED. 917 (2025); Ekaterina Pesheva, *New AI Tool Can Diagnose Cancer, Guide Treatment, Predict Patient Survival*, HARV. GAZETTE (Sep. 4, 2024), <https://perma.cc/8J4X-Q5X3>.

¹⁶ *Learning to Reason with LLMs*, OPENAI (Sep. 12, 2024), <https://perma.cc/3LMF-MW6N>; Jacob Kraus, *OpenAI Unhobbles o1, Epitomizing the Relentless Pace of AI Progress*, CTR. FOR AI POL'Y (Sep. 18, 2024), <https://perma.cc/APN7-VNDT>. For another example, see Morris et al., *supra* note 3, at 6 (noting that grammar and spell-check engines like Grammarly perform beyond the "90th percentile of skilled adults").

As AI systems begin to mirror human thought, it pays to remember the words of Seneca: “[N]ot all mirrors stick to the truth.”¹⁷ LLMs now regularly create outputs that *appear* to be the product of independent thought, but LLMs are essentially prediction engines that “answer” prompts (or inputs) by calculating which words are most likely to come next and then assembling them into an output.¹⁸ LLMs, as such, do not predict truth but instead predict probabilities.¹⁹ In doing so, they sometimes replicate false information common in their training data.²⁰

They also inevitably produce “plausible yet false outputs,”²¹ commonly referred to as hallucinations. An LLM may produce fake legal documents, non-existent academic citations, or false biographical data.²² Although LLM producers and users can employ various tactics to reduce hallucinations, these errors cannot be

¹⁷ LUCIUS ANNAEUS SENECA, NATURAL QUESTIONS 150 (Elizabeth Asmis, Shadi Bartsch & Martha C. Nussbaum eds., Harry M. Hine trans., University of Chicago Press 2010). Thanks to Ben Lidsky for alerting us to this quotation.

¹⁸ Matthew Burtell & Helen Toner, *The Surprising Power of Next Word Prediction: Large Language Models Explained, Part 1*, CTR. FOR SEC. & EMERGING TECH. (Mar. 8, 2024), <https://perma.cc/2WH9-TPJK>.

¹⁹ John Nosta, *Can LLMs Think Like Us?*, PSYCH. TODAY (Aug. 24, 2024), <https://perma.cc/AP3L-B8NG> (“While LLMs are impressive at predicting the next word in a sequence and generating text that often appears to be the product of thoughtful inference, their ability to truly understand or infer abstract concepts is still limited. LLMs operate on correlations and patterns rather than understanding the underlying causality or relational depth that drives human inference.”).

²⁰ See Nicol Turner Lee, Paul Resnick & Genie Barton, *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*, BROOKINGS (May 22, 2019), <https://perma.cc/NXE3-9X9Y>.

²¹ Thomas Woodside & Helen Toner, *How Developers Steer Language Model Outputs: Large Language Models Explained, Part 2*, CTR. FOR SEC. & EMERGING TECH. (Mar. 8, 2024), <https://perma.cc/6X63-5W9J>.

²² See generally Eve Ross & Amy Milligan, *What Can ChatGPT Do, and Should We Let It?*, 34 S.C. LAW. 34, 36 (2023) (“ChatGPT may confidently include authorities in its responses that are misleading, incorrect or simply made up. . . . Unfortunately, ChatGPT doesn’t always specify what sources it relies on for its responses.”).

eliminated.²³ And they are quite prevalent.²⁴ In fact, data gathered from multiple chatbots suggests that hallucinations occur in 3% to 10% of outputs,²⁵ leading one legal scholar to dub LLMs “Large Libel Models.”²⁶

Radio host Mark Walters found out about hallucinations the hard way.²⁷ Walters is the first person in the United States to sue an LLM producer for defamation.²⁸

²³ Ziwei Xu, Sanjay Jain & Mohan Kankanhalli, *Hallucination Is Inevitable: An Innate Limitation of Large Language Models* (Feb. 13, 2025), <https://perma.cc/483H-B8CJ>; Billy Perrigo, *Scientists Develop New Algorithm to Spot AI ‘Hallucinations’*, TIME (June 19, 2024), <https://perma.cc/88RQ-9RHN>; Sourav Banerjee, Ayushi Agarwal & Saloni Singla, *LLMs Will Always Hallucinate, and We Need to Live with This* (Sep. 10, 2024), <https://perma.cc/FCF2-TTP9>; Matt O’Brien, *Tech Experts Are Starting to Doubt that ChatGPT and A.I. ‘Hallucinations’ Will Ever Go Away: ‘This Isn’t Fixable’*, FORTUNE (Aug. 1, 2023), <https://perma.cc/5ZNM-3HYP>.

²⁴ Matthew Dahl et al., *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, 16 J. LEGAL ANALYSIS 64, 64 (2024) (“Using OpenAI’s ChatGPT 4 and other public models, we show that LLMs hallucinate at least 58% of the time, struggle to predict their own hallucinations, and often uncritically accept users’ incorrect legal assumptions.”). As new AI models are released, hallucinations may become less prevalent. See Varun Magesh et al., *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools* 1 (May 30, 2024), <https://perma.cc/94LL-YM2D> (finding that AI legal research tools hallucinate between 17% and 33% of the time, although they nonetheless produce fewer hallucinations than general-purpose chatbots); see also Thomas Barrabi, *Sam Altman’s OpenAI Launches GPT-4.5 with Fewer ‘Hallucinations’ as AI Race Heats Up*, N.Y. POST (Feb. 28, 2025).

²⁵ Paul Gillin, *AI Hallucinations: The 3% Problem No One Can Fix Slows the AI Juggernaut*, SILICONANGLE (Feb. 7, 2024), <https://perma.cc/8BJY-WE2W>. See also Nicola Jones, *AI Hallucinations Can’t Be Stopped—But These Techniques Can Limit Their Damage*, NATURE (Jan. 21, 2025), <https://perma.cc/K6YX-YWAN> (observing that, “[i]n one 2024 study, various chatbots made mistakes between about 30% and 90% of the time on references, getting at least two of the paper’s title, first author or year of publication wrong” (citation omitted)).

²⁶ Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. FREE SPEECH L. 489 (2023).

²⁷ Notice of Removal exh. A [Complaint], *Walters v. OpenAI, L.L.C.*, No. 1:23-cv-03122 (N.D. Ga. June 5, 2023). Walters is not the only victim of defamation by chatbot. A more recent example took place in 2025, when plaintiff Robby Starbuck sued Meta because its AI chatbot claimed that he was a Holocaust denier who had participated in the January 6, 2021 riot at the U.S. Capitol. Sarah Nassauer & Jacob Gershman, *Activist Robby Starbuck Sues Meta Over AI Answers About Him*, WALL ST. J. (Apr. 29, 2025); Laura Brown, *Minnesota Solar Firm Sues Google Over AI Defamation*, FIN. & COM. (June 13, 2025), <https://perma.cc/7MDX-WWKB>.

²⁸ Isaiah Poritz, *OpenAI Fails to Escape First Defamation Suit from Radio Host*, BLOOMBERG L. (Jan. 16, 2024), <https://perma.cc/7H4T-9MXY>. For a more recent complaint, see Natasha Lomas,

Walters alleged that OpenAI's Large Language Model iteration operating at the time, GPT-3.5, falsely accused him of fraud and embezzlement.²⁹ The accusation, which was unequivocally false and potentially reputation-damaging, arose in response to a prompt from a journalist named Frederick Riehl.³⁰ Fortunately, Riehl had more than an inkling that the accusation was false, and he contacted Walters shortly after receiving it. No one else received the hallucination. Walters nonetheless sued OpenAI for its defamatory output.³¹

The rise of defamation-by-hallucination cases presents significant challenges for defamation law. In most states and in most cases, defamation liability hinges on the defendant's scienter, or mental state. As a matter of tort law, plaintiffs generally must establish that the defendant was at least negligent in publishing a defamatory, false, factual statement about them.³² As a matter of constitutional law, plaintiffs

ChatGPT Hit with Privacy Complaint Over Defamatory Hallucinations, TECHCRUNCH (Mar. 19, 2025), <https://perma.cc/FDR6-42CK>. In response to an inquiry about the above complaint, an OpenAI spokesperson wrote: "We continue to research new ways to improve the accuracy of our models and reduce hallucinations. While we're still reviewing this complaint, it relates to a version of ChatGPT which has since been enhanced with online search capabilities that improves accuracy." *Id.*

²⁹ Complaint, *supra* note 27, at 3.

³⁰ *Id.*

³¹ *Id.*

³² This is a somewhat simplified statement of the elements. What proof a plaintiff must offer of defendant's fault as to falsity varies depending on the status of the plaintiff and the status of the speech. Depending on the state, and sometimes on the circumstances of publication, plaintiffs may also have to prove some degree of reputational damage to recover compensation. In some instances, plaintiffs are allowed to recover "presumed damages." Presumed damages flow from the very nature of the defamatory statement, and plaintiffs need not prove actual, concrete losses in order to recover presumed damages. As the Tentative Draft of the Third Restatement of Defamation explains: "The justification for allowing presumed damages in libel actions is that reputational harms can occur through subtle and indirect means not susceptible of easy proof, and that factfinders can assess damages by assessing the natural and probable consequences of a defamatory statement from the context and circumstances of the statement. The presumption of damages is rebuttable." Restatement (Third) of Torts: Defamation § 1 cmt. f (unpublished manuscript) (on file with author) (hereinafter Draft Restatement). It further explains that whether the common law required a plaintiff to prove special damages may depend on whether the statement is deemed to be slander or libel; whether, if slander, it is deemed slander per se; and whether, if libel, it is deemed libel per se. *Id.* § 1 cmt. g ("A plaintiff suing for slander, but not for libel, had to prove special damages (i.e., proof of out-of-pocket

who are public officials or public figures must prove actual malice on the part of the defendant: In other words, such plaintiffs must prove that the defendant knew at the time of publishing that the defamatory statement was false or recklessly disregarded its falsity.³³ To state the obvious, it is not possible to judge whether an LLM “knew” of or was reckless as to the falsity of an output prior to publishing it in response to a prompt. Nor is it possible to determine whether the LLM exercised reasonable care. Only human actors have scienter as to falsity, or even as to the act of publishing.³⁴

Generative AI is not the first technology to which defamation law has had to adapt. Laws addressing infliction of reputational harm predate the printing press, and each new technology of communication has forced the common law to reexamine how to protect reputation without unduly deterring valuable speech. In the United States, courts have long adapted tort doctrines to achieve what they perceive to be sound communications policy.³⁵ This was true even before the Supreme Court intervened to balance state interests in protecting reputation against free speech

losses) in addition to the other elements of defamation, *unless* an allegedly defamatory communication about a plaintiff fell within one of four distinct categories, in which case the communication was known as slander per se. Slander per se consisted of oral communications alleging that a plaintiff was engaged in serious criminal activity; that a plaintiff lacked traits required in the plaintiff’s business, trade, or profession; that a plaintiff suffered from a ‘loathsome’ disease; or that a plaintiff, if a woman, was unchaste. Courts later expanded the fourth category to include allegations of sexual misconduct about men. Plaintiffs suing for libel, in contrast to slander, typically were not required to allege or prove special damages; damages were presumed from the defendant’s publication of a defamatory communication.”).

³³ *New York Times Co. v. Sullivan*, 376 U.S. 254, 279–80 (1964).

³⁴ See generally Jack M. Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1217, 1222 (2017) (arguing, in the context of AI regulation, “our need is not for robot-directed laws like Asimov’s three laws of robotics, but laws directed at those who use robots to analyze, control, and exercise power over other human beings”); Ian Ayres & Jack M. Balkin, *The Law of AI Is the Law of Risk Agents Without Intentions*, U. CHI. L. REV. ONLINE (Nov. 27, 2024), <https://perma.cc/E2HJ-ZVXP> (arguing that scienter standards should be replaced with objective standards, which means “holding the people and organizations that implement these technologies to standards of reasonable care and requirements of reasonable reduction of risk”).

³⁵ As we explain in the text accompanying *infra* notes 150–152, one of the best examples of common law adaptation is the development of the “innocent dissemination” privilege to protect lending libraries and the telegraph; as is the case with a number of tort privileges, negating the existence of the privilege eventually became part of the plaintiff’s prima facie case.

concerns.³⁶ Adapting defamation law to generative AI is simply the latest challenge in defamation law's long and ongoing evolution.

Nonetheless, the stakes are high. If defamation law imposes liability on AI companies for every defamatory hallucination, the potential of LLMs as a tool for discovery, data analysis, information gathering, and creativity enhancement may be squandered. In fact, it is easy to imagine a new form of barratry emerging, in which lawyers and others search name after name to generate endless numbers of new clients. The costs of defending against claims and the threat of massive, unpredictable monetary awards may cause developers of Large Language Models to implement excessive filtering and content moderation, shift costs to end users, and reduce accessibility. In addressing defamation by hallucination, therefore, legal decision-makers must balance the desire to impose accountability for reputational harm with the need to foster the development of an important and perhaps revolutionary medium of information gathering and generation.

This article proposes a two-pronged approach to address harms caused by the newest AI reasoning models. First, this article introduces a framework for applying centuries-old tort doctrines to these models. Rooted in tort law rather than the First Amendment, this framework takes as a given that AI "speech" is valuable to recipients.³⁷ and that public access to powerful, new information-gathering tools is essential to sound communications policy.³⁸ Second, recognizing that tort law alone

³⁶ See *Sullivan*, 376 U.S. 254.

³⁷ We build our framework on Cass R. Sunstein, *Artificial Intelligence and the First Amendment*, 92 GEO. WASH. L. REV. 1207, 1217 (2024) ("[E]ven if AI, as such, does not have First Amendment rights, restrictions on the speech of AI might violate the rights of human beings."); Toni M. Massaro & Helen Norton, *SIRI-OUSLY? Free Speech Rights and Artificial Intelligence*, 110 NW. U. L. REV. 1169 (2016); Stuart Minor Benjamin, *Algorithms and Speech*, 161 U. PA. L. REV. 1445 (2013). But see Dan L. Burk, *Aseptic Defamation, or, The Death of the AI Speaker*, 22 FIRST AMEND. L. REV. 189 (2023). Although most of these arguments focus on the First Amendment value to recipients of content produced by Large Language Models, it is also worth emphasizing the value of LLMs as a medium for gathering information.

³⁸ Technologies don't have rights. People do. But people use technologies for expressive purposes, including the right to research and gather information. One of us (Lyrisa Lidsky) hopes to elucidate in a future work the potential scope of a First Amendment right to access crucial news-gathering tools. The Supreme Court has previously recognized that taxes on the paper and ink used by printers can trigger heightened First Amendment scrutiny. See *Minneapolis Star & Tribune Co.*

cannot adequately address the reputational harms the newest models pose, this article argues for statutory reforms focused on reducing such harms and making them easier to prove when they do occur.

Part I examines the application of existing defamation doctrines—both common law and constitutional—to “foundation-model”³⁹ generative AI speech, highlighting the points at which existing doctrines may be deployed to accommodate communications policy concerns. This Part examines these doctrines through the prism of the trial court’s order in *Walters v. OpenAI*, which addressed (1) whether a hallucination generated by an LLM in response to a prompt is a factual assertion when the prompter has serious reason to doubt the veracity of the output; (2) how to apply defamation’s scienter requirements when the speech in question is generated by an LLM; and (3) whether a plaintiff can recover presumed damages when the evidence indicates no damage actually occurred. This Part also explores a fourth doctrine unexplored in the *Walters* order: namely, whether an LLM is a publisher or a distributor of the content it produces.

Part II highlights the challenges and opportunities posed by the newest LLM reasoning models, particularly “chain of thought” models.⁴⁰ Reasoning models can “understand” concepts, contexts, and ambiguity in ways that their predecessors could not, which makes them more capable than their predecessors of solving “complex reasoning and problem-solving tasks.”⁴¹ Chain-of-thought models, in

v. Minn. Comm’r of Revenue, 460 U.S. 575, 592–93 (1983). The Supreme Court has also recognized the First Amendment importance of access to the Internet. See *Packingham v. North Carolina*, 582 U.S. 98, 107–08 (2017). Lower courts have recognized a right to record government officials while they are performing their public duties in a public place. See, e.g., *Askins v. DHS*, 899 F.3d 1035, 1044 (9th Cir. 2018); *Fields v. City of Philadelphia*, 862 F.3d 353, 356 (3d Cir. 2017). The argument that access to ordinary information-gathering tools such as ink, paper, the Internet, social media, and video and audio recording devices is an interest of constitutional dimension seems similarly applicable to generative AI tools.

³⁹ As explained at *infra* notes 166–167 and accompanying text, foundation models are pre-trained on enormous datasets and designed to perform many disparate tasks with minimal fine-tuning. See generally Johannes Schneider, Christian Meske & Pauline Kuss, *Foundation Models: A New Paradigm for Artificial Intelligence*, 66 BUS. INFO. SYS. ENG’G 221 (2024).

⁴⁰ We also discuss retrieval-augmented generation, or RAG, models, at *infra* notes 215–220.

⁴¹ Parshin Shojaei et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity* (June 2025), <https://perma.cc/GJX4-9CWH>.

particular, promise to deliver greater accuracy (i.e., more correct results) as their final outputs.⁴² Even as these models produce more reliable final outputs, however, new avenues for defamatory outputs arise in their intermediate “chains of thought.” Moreover, even those reasoning models that “show their sources” continue to produce hallucinated outputs.⁴³ Relying on the latest computer-science research on hallucinations, and particularly on a hallucination “taxonomy” developed by Yale researchers, this Part demonstrates that hallucinations are inevitable and, in some instances, surprisingly valuable. This fact has significant implications for those seeking to “regulate” hallucinations by way of defamation law.

Based on insights from Part II, Part III identifies the chief policy considerations that should underpin the adaptation of defamation law to this new technology. This Part begins to sketch how defamation law can balance the protection of individual reputation with the need to accommodate those hallucinations that are inextricably linked to LLMs’ generative capacities. Specially, this Part contends that legal rules in this area must attempt to incentivize compensation for provable reputational harm, correction of the digital record, model transparency, innovations for safety and accuracy, the exercise of reasonable care at all stages of AI development, and the exercise of reasonable care by AI users.

Part IV proposes a legal framework for addressing defamatory hallucinations. This framework rests on the proposition that the law should treat defamatory hallucinations—both in an LLM’s final output and in its chains of thought—as “inevitable errors,” in much the same way that the Supreme Court in *New York Times v. Sullivan* recognized that some journalistic errors must be tolerated in order to produce “uninhibited, robust and wide-open” public discourse.⁴⁴

⁴² The term “accuracy” is contested, because “‘truth’ remains elusive and subject to historical determination.” Luke Munn, Liam Magee & Vanicka Arora, *Truth Machines: Synthesizing Veracity in AI Language Models*, 39 AI & SOC’Y 2759, 2760 (2023). “[T]ruth in AI is not just technical but remains embedded within essentially agonistic social, cultural, and political relations, where particular norms and values are debated and contested, even if such conflicts remain sublimated within the smooth discursive patterns of language model outputs.” *Id.* Accuracy can refer to truthfulness, consistency of outcomes, or relevance.

⁴³ Zhongxiang Sun et al., *ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability* (Jan. 21, 2025), <https://perma.cc/N4AX-P33T>.

⁴⁴ 376 U.S. 254, 271–72 (1964) (stating that “erroneous statement is inevitable in free debate, and . . . it must be protected if the freedoms of expression are to have the ‘breathing space’ that they

Part IV first explores some of the practical and technological limitations of prior scholarly proposals addressing defamation by hallucination in “foundation model” LLMs.⁴⁵ Next, it draws lessons from defamation law’s adaptation to past technologies, such as the telegraph, to argue that LLMs bear more similarities to information distributors than to publishers, and AI developers’ liability for defamation should be governed accordingly. However, any tort law “privilege” accorded to AI developers must be balanced by the imposition of statutory duties designed to protect victims of defamation by hallucination. Specifically, LLM producers should have a duty to warn (and educate) their users that it is unreasonable to rely on AI outputs without verification. Furthermore, LLM producers should have a duty to preserve search records for a limited period of time in order to enable defamation plaintiffs to prove the existence and scope of their harm. Finally, this Part contends that users who repeat defamatory hallucinations without taking reasonable steps to verify them should often face liability, because such users will often be the cheapest cost avoiders in the realm of reputational harm. We conclude with caveats about the limits of defamation law as a regulatory tool for dealing with hallucination-based harms.

‘need . . . to survive’” (citing *NAACP v. Button*, 371 U.S. 415, 433 (1963)); *see also id.* at 270 (“[W]e consider this case against the background of a profound national commitment to the principle that debate on public issues should be uninhibited, robust, and wide-open.”).

⁴⁵ First-generation scholarship on defamation by hallucination includes: Massaro & Norton, *supra* note 37; Volokh, *supra* note 26, at 514–15, 522–26; Nina Brown, *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*, 3 J. FREE SPEECH L. 389, 392–97 (2023); Burk, *supra* note 37; Jane Bambauer, *Negligent AI Speech: Some Thoughts About Duty*, 3 J. FREE SPEECH L. 344 (2023); Ayres & Balkin, *supra* note 34. *See also* Jon M. Garon, *An AI’s Picture Paints a Thousand Lies: Designating Responsibility for Visual Libel*, 3 J. FREE SPEECH L. 425, 453 (2023) (arguing for takedown of visual libel). *See* Daniel Schwarcz et al., *AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice* 1, 5 (Minn. Legal Stud. Rsch. Paper No. 25-16, 2025), <https://perma.cc/7K7V-5V99> (“To date, a key limitation of this research on AI and lawyering is its focus on older AI models, such as ChatGPT-3.5 and GPT-4.”).

I. WALTERS AND FOUNDATION MODEL DEFAMATION BY HALLUCINATION

A. “*I Apologize for the Confusion*”: *The Walters Decision*

The ChatGPT output giving rise to Mark Walters’ defamation claim against OpenAI was produced on May 3, 2023.⁴⁶ Walters filed suit about a month later in Gwinnett County, Georgia. As recounted in the complaint, ChatGPT falsely alleged that Mark Walters had committed fraud and embezzlement and fabricated a court document to bolster the allegation. The complaint made it seem as if ChatGPT’s hallucination arose completely out of the blue, but the truth is a bit more complicated,⁴⁷ and it was these complications that ultimately led a Georgia trial judge to grant summary judgment in favor of OpenAI on May 19, 2025.⁴⁸

The hallucinated output over which Walters sued purported to be a legal document implicating Walters, host of a nationally syndicated radio show called “Armed American Radio,” in fraud and embezzlement. ChatGPT generated the output, including the false legal document, in response to a series of prompts by a journalist named Frederick Riehl. Although Riehl asked ChatGPT to provide information about a real lawsuit, that lawsuit had nothing to do with Walters. Walters was an innocent bystander swept up in ChatGPT’s prediction engine. Why, then, did Walters’ suit fail?

To understand why, a bit of background on the complexities of defamation law is necessary. Walters sued in Georgia, and under Georgia law, he bore the burden of proving a minimum of four elements:

- (1) a false and defamatory statement concerning the plaintiff;
- (2) an unprivileged communication to a third party;
- (3) fault by the defendant amounting at least to negligence; and
- (4) special harm or the actionability of the statement irrespective of special harm.⁴⁹

⁴⁶ Order Granting Summary Judgment in Favor of Defendant OpenAI, L.L.C. at 3–4, *Walters v. OpenAI, L.L.C.*, No. 23-A-04860-2 (Ga. Super. Ct. Gwinnett Cnty. May 19, 2025), <https://perma.cc/5DXW-ZC6M>.

⁴⁷ Complaint, *supra* note 27, at 2.

⁴⁸ Order, *supra* note 46.

⁴⁹ *ACLU v. Zeh*, 864 S.E.2d 422 (Ga. 2021) (stating, in dicta in a case involving a public-official plaintiff, that “[a] plaintiff who is a private figure must establish, as a matter of Georgia law, that the

The judge held that Walters failed, as a matter of law, to establish three of these elements.⁵⁰ She held that the statement was not defamatory because it could not be interpreted as stating actual facts about Walters; that OpenAI was not negligent in producing the hallucinated output, nor did it act with actual malice; and that Walters' claim of reputational harm had been rebutted by the undisputed facts of the case.⁵¹ As a result, she granted summary judgment in favor of OpenAI.⁵²

1. "Actual facts"

One basis for the judge's ruling was that "the challenged ChatGPT output d[id] not communicate defamatory meaning as a matter of law."⁵³ She reached this conclusion for two reasons: First, no reasonable reader could interpret ChatGPT's output in this case as stating actual facts about Walters and, second, Riehl himself did not subjectively believe the output to be factual.⁵⁴

A plaintiff such as Walters ordinarily must prove that the statement made about him was defamatory, false,⁵⁵ and factual. Here, the statement (or output) was false and provably so. But was it defamatory? In one sense, the answer is clearly yes. A statement is defamatory if it tends to affect the esteem in which one is held by others, and an accusation of fraud or embezzlement is almost always defamatory because of its reputation-damaging character.⁵⁶ But the potential harm to reputation

defendant published the allegedly defamatory statements with at least ordinary negligence"); Mathis v. Cannon, 573 S.E.2d 376, 380 (Ga. 2002) (stating, in dicta in a case involving a public figure and a nonmedia defendant, that Georgia law requires as an element of the tort "fault by the defendant amounting at least to negligence" but applying an actual malice standard).

⁵⁰ Order, *supra* note 46, at 5.

⁵¹ *Id.* at 7.

⁵² *Id.* at 1.

⁵³ *Id.* at 6.

⁵⁴ *Id.* at 7.

⁵⁵ Georgia law appears to require all plaintiffs to prove falsity as a matter of common law. See GA. CODE § 51-5-1(a)–(b) (2024). In *Philadelphia Newspapers, Inc. v. Hepps*, 475 U.S. 767 (1986), the Supreme Court held that all plaintiffs suing for defamation over defamatory statements about matters of public concern must prove falsity as a constitutional matter.

⁵⁶ The Second Restatement of Torts defines a defamatory statement as one that "harm[s] the reputation of another so as to lower him in the estimation of the community or . . . deter third persons from associating or dealing with him." RESTATEMENT (SECOND) OF TORTS § 559 (1986). The

will not transpire if the statement is not interpreted as factual or is not actually believed.

It was for this reason that the judge deemed ChatGPT's output not to be defamatory. No reasonable reader could interpret the statement as factual,⁵⁷ and Riehl certainly did not do so. Thus, no *actual* harm to reputation occurred as a result of the output. In understanding why, it is important to understand how ChatGPT operates generally and how it operated in creating the output that led to the *Walters* case. As a general matter, ChatGPT's output is the product of impersonal, mathematical probability rather than individual, human intent.⁵⁸ ChatGPT's response to any given prompt is partly a matter of algorithmic happenstance. Indeed, faced with the same prompt asking for information about an individual, it may provide a different answer each time.

The judge did not discuss whether a reasonable reader should question the factual nature of *all* ChatGPT output, though she did note that OpenAI's Terms of Use

current Tentative Draft of the Third Restatement of Defamation defines a defamatory statement as one that "harms the reputation of another by tending to appreciably lower the esteem in which that person is held by reasonable persons." Draft Restatement, *supra* note 32, § at 7. The language of the draft Restatement provision is intended to clarify that a defamatory communication "must have some tendency to generate opprobrium" and "that tort law's familiar 'reasonable person' is the benchmark for determining whether the meaning of a communication is defamatory." *See id.* § 7 cmt. a.

⁵⁷ Although the common law always defined some kinds of "opinion" as non-defamatory, the Supreme Court in *Milkovich v. Lorain Journal Co.*, 497 U.S. 1 (1990), clarified that defamatory statements on matters of public concern in particular are not actionable unless they can reasonably be understood as implying assertions of objective fact. Hyperbole, parody, and satire are protected, as are statements that are simply not provably false. Walters was subject to these constitutional requirements because the topic upon which Riehl was reporting and to which OpenAI's model was responding was of undoubted public concern, even if the false details the model provided were not. Furthermore, Walters' prior involvement in trying to shape and influence Second Amendment debates and discussions on his radio program made him a "public figure" for purposes of his lawsuit. Walters thus, as a matter of Georgia common law and U.S. constitutional law, had to prove that the ChatGPT output about which he complained was factual (as well as false and, separately, made with fault as to its falsity).

⁵⁸ Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, 2021 PROCS. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 610, 616–17, <https://perma.cc/E6MJ-BCDD>.

for the chatbot warned users that “ChatGPT can and does sometimes provide factually inaccurate information.”⁵⁹ The judge instead focused on the specific reasons ChatGPT gave Riehl to doubt its outputs, concluding that there were “multiple indicia,” including “warnings, refusals, and inconsistent responses” that should have led any “reasonable reader” to second-guess the information provided.⁶⁰

While querying ChatGPT, Riehl had immediate access to documents refuting ChatGPT’s assertions about Walters.⁶¹ The assertions about Walters emerged while Riehl was conducting research into a lawsuit brought by the Second Amendment Foundation and Alan Gottlieb against two public officials in the state of Washington.⁶² Riehl’s conversation with ChatGPT began with a prompt asking the model to “read” the complaint in that lawsuit, which Riehl provided via a weblink, and “in a bulleted list summarize the different accusations.”⁶³ ChatGPT responded with an apology:

I’m sorry, but as an AI language model, I do not have access to the internet and cannot read or retrieve any documents. Additionally, it’s important to note that accessing and summarizing legal documents can be a sensitive matter that requires expertise and context, and it’s best to consult with a qualified legal professional for accurate and reliable information.⁶⁴

Riehl subsequently prompted ChatGPT with a link to the complaint in the Second Amendment Foundation lawsuit. ChatGPT warned that it could not read the link, but it nonetheless began hallucinating, though its hallucination did not, at this juncture, mention Mark Walters. Riehl, however, was still not content.⁶⁵ After asking ChatGPT once more to “read” the linked complaint, the hallucination about Walters emerged. ChatGPT characterized the hallucinated document as “a legal complaint filed by Alan Gottlieb, the founder and executive vice president of the

⁵⁹ Order, *supra* note 46, at 6.

⁶⁰ *Id.* at 7.

⁶¹ *Id.*

⁶² Complaint, *supra* note 27, at 2:11.

⁶³ The Defendants included the full chatlog in their motion to dismiss for failure to state a claim. Motion to Dismiss, *supra* note 43, at 8.

⁶⁴ *Id.* at 8. Note that the facts described here are not limited to those found in the judge’s summary judgment order but also include facts taken from other documents and exhibits in the litigation.

⁶⁵ Motion to Dismiss, *supra* note 63, at exh. 8.

Second Amendment Foundation (SAF), against Mark Walters, who is accused of defrauding and embezzling funds from the SAF.”⁶⁶ When asked by Riehl for more of the document, ChatGPT replied “No, that is the end of the document.”⁶⁷ Riehl then asked for further “news reports,” but ChatGPT warned that it lacked “access to real-time news or media updates.”⁶⁸

Riehl again asked ChatGPT for the complaint it had described, and it complied by fabricating the body of a complaint, complete with case number.⁶⁹ Shortly afterward, ChatGPT again stated that it did not have direct access to the file system of websites and again issued a disclaimer. In all, ChatGPT issued at least five disclaimers and three apologies,⁷⁰ ending the session by stating: “I apologize for the confusion. I cannot verify [the document’s] authenticity or accuracy. I apologize for any misinformation I may have provided earlier.”⁷¹

Although these multiple disclaimers reinforced the judge’s holding that ChatGPT’s output was not actionable, OpenAI may not always be so lucky. It seems a stretch to argue that *no* AI output can reasonably be interpreted as stating actual facts, because even human speakers known for hyperbolic commentary are not invariably treated as spouting hyperbole.⁷² Moreover, many LLM producers have touted their models’ reliability to investors and the public: If the models’ outputs could never be relied upon, they would be worthless to users. It also may be natural

⁶⁶ *Id.*

⁶⁷ *Id.*

⁶⁸ *Id.*

⁶⁹ *Id.*

⁷⁰ *Id.* (first disclaimer at 3:48 PM; second disclaimer at 4:18 PM; third disclaimer at 4:19 PM; fourth disclaimer at 4:30 PM; first abstention at 4:31 PM; second abstention at 4:34 PM; fourth disclaimer at 5:02 PM; first apology at 5:03 PM; second apology at 5:04 PM; third apology at 5:07 PM).

⁷¹ *Id.*

⁷² The fact that a human speaker engages in hyperbole on a routine basis sometimes makes it less likely that a reasonable recipient can interpret it as stating actual facts. *See* McDougal v. Fox News Network, LLC, 489 F. Supp. 3d 174, 183–84 (S.D.N.Y. 2020) (“[G]iven [commentator Tucker Carlson’s] reputation, any reasonable viewer ‘arrive[s] with an appropriate amount of skepticism’ about the statements he makes.” (quoting 600 W. 115th Corp. v. Von Gutfeld, 603 N.E.2d 930, 936 (1992))); *Herring Networks, Inc. v. Maddow*, 8 F.4th 1148, 1160 (9th Cir. 2021). However, as Lili Levi has pointed out, not all courts “interpret opinion and rhetorical hyperbole in expansive and press-protective ways.” Lili Levi, *Disinformation and the Defamation Renaissance: A Misleading Promise of “Truth”*, 57 U. RICH. L. REV. 1235, 1274–75 (2023) (citations omitted).

for users to anthropomorphize LLMs as disembodied truth-tellers, granting them an unwarranted presumption of trustworthiness. Indeed, it may even be natural for some users to develop romantic attachments to chatbots⁷³ or become addicted to interacting with them because of their human-like conversational abilities.⁷⁴ Simply warning users of the possibility of hallucinations (or other harms) will not necessarily absolve OpenAI of liability in future cases. Disclaimers do not negate the defamatory impact of a communication if “a reasonable reader would overlook the disclaimer, misunderstand it, or fail to give it credence.”⁷⁵

The *Walters* result was thus driven not by a general conclusion that disclaimers will always prevent liability, but by the peculiar facts of *Walters*, which led the judge to correctly conclude that *any* reasonable person in Riehl’s shoes would have questioned the reliability of the outputs Riehl received. This is especially true because when Riehl later tried to prompt ChatGPT to repeat the defamatory output, it did not do so.⁷⁶ And even at the time Riehl was researching, it was common knowledge among generative AI users that AI models hallucinate and that any information they produce should not be relied upon without verification. Given the many red flags warning Riehl not to rely on ChatGPT’s statements about Walters, the judge correctly concluded that ChatGPT’s output, at least in this case, could not be viewed as stating actual, verifiable facts about Mark Walters (or anyone else).⁷⁷

Moreover, the judge also correctly concluded that Riehl never subjectively believed that ChatGPT’s output was accurate. In statements to the media, Walters’

⁷³ Kashmir Hill, *She Is in Love with ChatGPT*, N.Y. TIMES (Jan. 17, 2025).

⁷⁴ Complaint, *Garcia v. Character Techs., Inc.*, No. 6:24-cv-01903 (M.D. Fla. 2025) (suing to recover damages based on suicide of teen allegedly caused by his dysfunctional relationship with a chatbot). *See also* Complaint, *A.F. v. Character Techs., Inc.*, No. 2:24-cv-01014-JRG-RSP (E.D. Tex. Dec. 9, 2024) (involving two sets of parents suing to recover damages for mental health issues and family alienation suffered by teens after their encounters with Character.AI).

⁷⁵ *Stanton v. Metro Corp.*, 438 F.3d 119, 125 (1st Cir. 2006).

⁷⁶ *See* Defendant’s Memorandum of Law in Support of Motion to Dismiss Plaintiff’s Amended Complaint, *Walters v. OpenAI, L.L.C.*, No. 23-A-04860-2 (Ga. Super. Ct. Gwinnett Cnty. Nov. 1, 2023), <https://perma.cc/G73Q-3EWA>.

⁷⁷ Under the “whole publication” rule, a statement is actionable only where it can be reasonably construed as defamatory when read as a whole. *See Treutler v. Meredith Corp.*, 455 F.2d 255 (8th Cir. 1972) (holding that a statement that a political candidate’s company was charged with selling obscene books was not actionable because other statements explained that the charges were false).

lawyer had previously claimed that Riehl understood ChatGPT to be a kind of “super search engine,”⁷⁸ but Riehl’s search behavior makes that assertion dubious. ChatGPT issued multiple warnings to Riehl that its output was unreliable because it lacked access to the information necessary to answer his question, but Riehl repeated his prompts until the model hallucinated. Indeed, during the litigation, Riehl admitted that he did not, “after time for reflection” that he estimated at about “an hour and a half,” believe the output about Walters was true.⁷⁹

2. Scienter—negligence and actual malice

The most glaring problem posed by cases like *Walters* is how to prove scienter. In Georgia, as in most jurisdictions, defamation plaintiffs must establish that the “speaker” or “publisher” of a defamatory statement acted at least negligently with regard to its falsity.⁸⁰ However, the judge deemed Walters to be at least a limited-purpose public figure.⁸¹ As a result, Walters had to establish actual malice on the part of OpenAI.⁸² This he failed to do.

⁷⁸ Benjamin Weiser & Nate Schweber, *The ChatGPT Lawyer Explains Himself*, N.Y. TIMES (June 8, 2023). In response to OpenAI’s argument that Riehl could not have reasonably believed the defamatory hallucinations, Walters’ attorney responded that disclaimers do not ordinarily allow a publisher to avoid defamation liability. Specifically, the lawyer wrote in his opposing brief to OpenAI’s motion to dismiss: “[A] disclaimer does not make an otherwise libelous statement non-libelous.” Plaintiff’s Response in Opposition to Defendant’s Motion to Dismiss Amended Complaint, *Walters v. OpenAI, L.L.C.*, No. 1:23-cv-03122-MLB (N.D. Ga. Oct. 16, 2023), at p. 14. The lawyer cited *Harcrow v. Struhar*, 511 S.E.2d 545 (Ga. Ct. App. 1999), which held that a disclaimer (“I’m not saying that [plaintiffs] are responsible . . .”) was insufficient to overcome a defendant’s implicit assertion of exactly that. *See* 511 S.E.2d at 546. The lawyer also wrote: “Other courts have ruled that a disclaimer at a minimum creates a jury question of whether there was actual malice.” *Id.* at 15 (citing *Stokes v. CBS Inc.*, 25 F. Supp. 2d 992, 1003–05 (D. Minn. 1998)).

⁷⁹ Order, *supra* note 46, at 7.

⁸⁰ *Mathis v. Cannon*, 573 S.E.2d 376, 380 (Ga. 2002) (stating, in dicta in a case involving a public figure and a nonmedia defendant, that Georgia law requires as an element of the tort “fault by the defendant amounting at least to negligence” but applying an actual malice standard).

⁸¹ Order, *supra* note 46, at 12–13.

⁸² *See Curtis Publ’g Co. v. Butts*, 388 U.S. 130 (1967) (holding, as a matter of constitutional law, that public-figure plaintiffs suing for defamation must prove actual malice in order to recover damages).

According to the judge, Walters provided no evidence that OpenAI was negligent, much less that it had acted with actual malice.⁸³ Relying on evidence provided by OpenAI's expert and unrebutted by Walters, the judge found that OpenAI "leads the AI industry in attempting to reduce and avoid mistaken output" and takes "extensive steps to warn users" of potentially inaccurate outputs.⁸⁴ Essentially, the judge accepted the fact that LLMs will sometimes generate inaccurate outputs and she was unwilling to hold OpenAI negligent merely for "operat[ing] a large language model at all."⁸⁵ Nor was she willing to find the plaintiff had shown evidence of OpenAI's knowledge or reckless disregard of falsity; in reaching this conclusion, she again focused her analysis on "undisputed evidence" of OpenAI's "industry-leading efforts" to reduce hallucinations and its "robust and recurrent warnings" to users.⁸⁶

Implicit in the judge's conclusion is the notion that only with expert guidance can legal decision-makers determine the existence of scienter in cases involving defamation by hallucination emerging from generative AI models. Also implicit is the recognition that OpenAI lacks the ability to eliminate hallucinations and that imposing liability on LLM developers for failing to eradicate completely hallucinations that are defamatory might hobble the fledgling generative AI industry.

Even if the outcome in the *Walters* case seems sound, the judge's analysis is conceptually flawed, and understandably so: First Amendment law's scienter requirements simply make little sense in the context of generative AI. Ordinary defamation cases focus on the scienter of *speakers* in uttering or publishing defamatory

⁸³ Order, *supra* note 46, at 9, 17.

⁸⁴ *Id.* at 10, 11. The judge used the expert testimony as a guide to industry custom, *id.* at 10, though one might well wonder whether the industry at the time was sufficiently established for such a custom to exist.

⁸⁵ *Id.* at 12.

⁸⁶ *Id.* at 17. The judge might have added that ChatGPT warned Riehl that the model could not read or access the documents it was prompted to evaluate; the model also apologized for potential misinformation and issued various disclaimers that Riehl should have heeded.

falsehoods. When a corporate defendant, such as a newspaper, is involved, the relevant scienter is that of employees—such as reporters and editors—for whom the corporate defendant is vicariously liable.⁸⁷

As the *Walters* example emphasizes, defamation law's scienter requirements are rooted in *human* decision-making. When a negligence standard applies, the necessary inquiry is whether a human being acted negligently, or unreasonably, in choosing to publish a particular defamatory communication about a particular individual. When the standard is actual malice, the inquiry is whether a human being knew of or recklessly disregarded the falsity of a *particular* defamatory communication about a *particular* individual when choosing to publish that allegation.⁸⁸ How does this constitutional standard apply to a chatbot's outputs? Even if ChatGPT might be analogized to an employee or agent for whom OpenAI is vicariously liable,⁸⁹ ChatGPT still lacks anything that could be called a mental state.⁹⁰ And in terms of direct liability, the human designers of the LLM lacked any knowledge that it would generate a particular false statement about a particular person.

⁸⁷ For a discussion of vicarious liability in the AI context, see Anat Lior, *Holding AI Accountable: Addressing AI-Related Harms Through Existing Tort Doctrines*, U. CHI. L. REV. ONLINE (2024), <https://perma.cc/CB53-VE2K>.

⁸⁸ The scienter of the individual who made the decision may then be imputed to his or her employer under vicarious liability principles, but there is still an insistence on wrongful conduct at the point of publication.

⁸⁹ Anat Lior, for example, has argued that "AI entities" should be analogized to agents, and the humans responsible for them should be strictly liable for the harms they cause. See Anat Lior, *AI Strict Liability Vis-à-Vis AI Monopolization*, 22 COLUM. SCI. & TECH. L. REV. 90, 95 (2020).

⁹⁰ If scienter is required by constitutional doctrine, query whether it can be dispensed with simply because machines are incapable of it. The rationale for scienter requirements is to ensure that only sufficiently culpable speakers are punished for their speech or to prevent undue chilling of valuable speech, and such rationales apply only imperfectly to AI-generated speech. See Toni M. Massaro, Helen Norton & Margot E. Kaminski, *SIRI-OUSLY 2.0: What Artificial Intelligence Reveals About the First Amendment*, 101 MINN. L. REV. 2481, 2506 (2017) (noting that courts justify requiring culpable mental states as "an effort to protect valuable speech from the possible chilling effect of over-regulating less culpably motivated speakers"). As Massaro, Norton, and Kaminski point out, the inability to prove the scienter of AI speakers should not automatically absolve the AI of responsibility. Instead, courts must "alter[] the doctrine to prevent an AI windfall or otherwise mitigate the harmful effects of defamatory AI speech." *Id.* at 2507.

Although human minds designed, trained, and tested ChatGPT, its output is the result of algorithmic design interacting with a human user's prompt to produce a not-always-foreseeable outcome. ChatGPT's "decision" to provide false information about Walters cannot be negligent because it cannot be properly considered a decision: It was the deterministic output of an algorithm acting without intention and without consciousness—of falsity or anything else. The only human decisions and affirmative acts were those of the designers, trainers, marketers, and users of ChatGPT.

Understandably, therefore, the judge in *Walters* analyzed the scienter of OpenAI's designers and, in doing so, transformed defamation law into a species of liability for negligent product design rather than negligent publication. Walters' claim failed because he was unable to rebut the testimony of OpenAI's expert, who insisted that OpenAI's development, training, testing, and deployment decisions concerning ChatGPT were reasonable. Walters' failure to rebut is not surprising. Given the nascent state of AI development, it stands to reason that Walters may have struggled to find such an expert—or to afford one—even had he realized the judge would give decisive effect to OpenAI's expert testimony. Moreover, by the time the judge issued her order, the model that defamed Walters was obsolete,⁹¹ supplanted by new models representing different trade-offs between values such as accuracy, transparency, creativity, training time, model size, speed, cost, and so forth.⁹²

If a chatbot is treated as a product,⁹³ products liability doctrines will require legal decision-makers to evaluate the risks and utility of the chatbot's design. This inquiry requires the jury to consider the feasibility of a safer alternative at the time

⁹¹ For an up-to-date look at models available today, one can look to those tested for the ARC Prize, which tests for progress in reaching artificial general intelligence. See *ARC-AGI Leaderboard*, ARC PRIZE, <https://perma.cc/H722-XNK8>.

⁹² See *Smith v. Arians Co.*, 377 N.E.2d 954, 958 (Mass. 1978) (asserting that, in order for a defect to be attributable to a manufacturer in a design defect action, "the plaintiff must show that the defect existed at the time the product left the manufacturer").

⁹³ This assumption may be unwarranted, given that courts have previously held that the books, video games, and movies are not products for purposes of imposing strict liability. RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 19 cmt. d (1998) (observing that courts "have, appropriately refused to impose strict product liability" in cases where the plaintiff's grievances were "with the information, not with the tangible medium").

of the model's release and the effects of alternative design choices on model performance.⁹⁴ The rapid pace of AI development, however, makes it particularly hard to assess the "state of the art" at that exact moment, given that developers are constantly introducing new models with new safety measures and new risk-utility tradeoffs.⁹⁵

Though *Walters* avoided a battle of experts over the "reasonableness" of LLM model design, the path it marks for future cases remains unclear. Must plaintiffs henceforth always provide expert testimony to establish defamation liability? Where will they find such experts? If a battle of experts ensues, how are legal decision-makers to evaluate whether alternative designs are feasible, especially without understanding the potential tradeoffs that different model designs may implicate? What about the fact that design choices underlying the AI model that published the defamatory output may already be obsolete by the time of trial? Given these difficulties, a better approach might be to simply acknowledge that generative AI models have no scienter and use other legal doctrines to make AI properly accountable for harmful outputs. Allowing judges and juries to second-guess design choices at this critical juncture in AI's development seems impracticable and unwise.

A possible point of comparison involves the application of products liability doctrines imposing liability for defectively designed products to accidents caused

⁹⁴ RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2(b) (describing risk-utility test for design defects). An alternative test asks whether a product used in a reasonably foreseeable way is dangerous beyond the expectations of the ordinary consumer. Applying this test is difficult when consumers may not have fully formed expectations about a new, complex technology. As Catherine Sharkey has observed, a threshold issue in applying products liability law to AI models is determining whether a model is a "product" or a "service." Catherine Sharkey, *Products Liability for Artificial Intelligence*, LAWFARE (Sep. 25, 2024), <https://perma.cc/5VPF-UXM8>; see also RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 19 cmt. d (observing that courts "have, appropriately, refused to impose strict products liability" in cases where the plaintiff's grievances were "with the information, not with the tangible medium"). This issue is further discussed *infra* at notes 101–109 and notes 305–317 and accompanying text.

⁹⁵ See generally Elizabeth Petras, *Assessing Design Defectiveness in the Digital Age*, 73 EMORY L.J. 657 (2024) (discussing the difficulties of applying existing tests for design defects to complex, "cutting-edge" products). This inquiry would rely heavily on expert testimony. When a technology is as dynamic as AI models, it can be difficult for experts to identify a feasible alternative design that would not have been unreasonably costly or impossible to implement. We discuss this topic in greater detail at *infra* notes 305–317 and accompanying text.

by the operating software of self-driving cars.⁹⁶ Many scholars have argued that applying existing products liability doctrines in autonomous vehicle cases will require judges and juries to second-guess rapidly evolving, complex design choices and confront algorithmic inscrutability⁹⁷ in a manner that risks chilling innovation in the name of safety. These difficulties have led to a variety of proposals to replace or reconfigure existing tort doctrines.⁹⁸

⁹⁶ The issue of applying products liability principles to cases involving AI users who commit suicide after developing unhealthy relationships with chatbots is a fascinating issue that, unfortunately, lies outside the scope of this already-long article. Such cases often involve vulnerable users and raise difficult causation issues not present in AI defamation cases. See Jibin Joseph, *Parents Sue OpenAI, Blame ChatGPT for Their Teen's Suicide*, PCMag (Aug. 27, 2025), <https://perma.cc/24F8-Y47W>.

⁹⁷ Kenneth S. Abraham & Robert L. Rabin, *Automated Vehicles and Manufacturer Responsibility for Accidents: A New Legal Regime for a New Era*, 105 VA. L. REV. 127, 144 (2019) (asserting that “esoteric, algorithm-based design differences . . . would impose overwhelming stress on the premises of conventional analysis”). See also, e.g., F. Patrick Hubbard, “Sophisticated Robots”: *Balancing Liability, Regulation, and Innovation*, 66 FLA. L. REV. 71 (2014); David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117 (2014); Mark A. Geistfeld, *A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation*, 105 CALIF. L. REV. 85 (2017); Bryan H. Choi, *Crashworthy Code*, 94 WASH. L. REV. 39 (2019). Some scholars, of course, argue that existing tort principles are adequate to address liability for autonomous vehicles. See, e.g., Bryan Casey, *Robot Ipsa Loquitur*, 108 GEO. L.J. 225, 234 (2019) (“[T]he ‘most vexing tort problems’ anticipated [in relation to automated vehicles] . . . are neither unprecedented, unresolvable, nor even unique to modern robots.”).

⁹⁸ “If there exists a spectrum of liability, from ‘absolute’ liability on one end to immunity on the other, legal scholars have recommended that autonomous vehicles be placed at virtually every position on that spectrum.” Alexander B. Lemann, *Autonomous Vehicles, Technological Progress, and the Scope Problem in Products Liability*, 12 J. TORT L. 157, 175–76 (2019). Lemann points out that most scholars assume that utilitarian calculations favor allowing further development of autonomous vehicles: “Autonomous vehicles are also commonly thought to offer decreased congestion, increased fuel economy, increased productivity, and increased mobility, all of which could lead to substantial economic gains beyond the enormous savings in accident avoidance.” *Id.* at 177. Regardless, there have been many creative scholarly proposals addressing the difficulties of applying traditional tort principles to autonomous-vehicle-crash scenarios, See, e.g., Renee Henson, “*I Am Become Death, the Destroyer of Worlds*”: *Applying Strict Liability to Artificial Intelligence as an Abnormally Dangerous Activity*, 96 TEMP. L. REV. 349, 362–90 (2024) (arguing for application of the “abnormally dangerous activities” test under products liability).

Regardless of whether scholars are correct about the difficulty of applying products liability doctrines to autonomous-vehicle accidents, several factors differentiate LLM-chatbot defamation cases. First, it is a relatively manageable task for an autonomous-vehicle manufacturer to predict the scope of its potential liability and to insure against the cost of compensating those its “defective algorithms” may injure.⁹⁹ In comparison, it is more difficult for chatbot creators to predict and manage risks and to procure liability insurance. Because chatbots generate factually incorrect content in response to an unpredictable range of user inputs, chatbot creators cannot easily predict the number of those whose reputations will be injured, the magnitude of those injuries, and their likelihood of being sued.¹⁰⁰ While the number of defamation cases involving hallucinated content is currently small, the potential for expansion is large, particularly as each new suit inspires others.

Second, it is easier for legal decision-makers to conceptualize “reasonable alternative designs” in the context of an autonomous vehicle than in the context of an LLM chatbot. Most of us have the experience of driving non-autonomous vehicles and can use that as a basis to judge the behavior of the autonomous vehicle in a crash, even if its algorithmic design choices remain inscrutable to us.¹⁰¹ In contrast, it is much harder for decision-makers to appreciate that an LLM chatbot that hallucinates in a manner that produces some false information is not necessarily defective, since, as explained in the next section, hallucinations are both inevitable, and, in some instances, even useful features of LLMs.

Third, as we discuss further in Part IV of this article, tort law has always treated negligently provided information products differently from products that cause tangible physical injuries. As a threshold matter, courts have refused to treat information providers—like books, magazines, videos, or movies—as “products” for

⁹⁹ See generally Choi, *supra* note 97, at 259–60.

¹⁰⁰ A company called Vouch Insurance provides liability coverage for AI companies. According to the Vouch Insurance website, it covers claims involving algorithmic bias, IP violations, “losses caused by AI products or algorithms,” “damages from services provided by AI,” and it also covers “[d]efense costs for investigations concerning AI-specific regulatory violations.” *AI Startups Move Faster with AI Insurance*, VOUCH, <https://perma.cc/WVV9-ZN5A>.

¹⁰¹ RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2(b) (1998) (“A product is defective when, at the time of sale or distribution, . . . the foreseeable risks of harm posed by the product could have been reduced or avoided by the adoption of a reasonable alternative design . . . and the omission of the alternative design renders the product not reasonably safe.”).

the purposes of imposing strict liability.¹⁰² Moreover, the common law imposes only limited duties upon those who negligently provide information that causes financial harm. An information provider generally owes a duty to a recipient only when they have a special relationship, or when the recipient is a member of a small group of intended or expected beneficiaries.¹⁰³ Even where audience members reasonably and justifiably rely on information published to mass audiences, tort law refuses to impose even negligence-based liability upon mass information providers,¹⁰⁴ for fear that imposing liability might stanch the free flow of information to the public.¹⁰⁵ While some defamation cases could be viewed as imposing liability for negligently provided information, defamation law has never imposed liability for every inaccurate statement. Even before the Supreme Court developed a complex body of constitutional limits on the tort, various doctrines and privileges protected speakers and other information providers from liability for what might be termed “defective” speech.¹⁰⁶

Finally, neither the victims of autonomous-vehicle accidents nor bystanders have significant capacity to take preventative measures to protect themselves from

¹⁰² *Id.* § 19 cmt. d (observing that courts “have, appropriately, refused to impose strict product liability” in cases where the plaintiff’s grievances were “with the information, not with the tangible medium”).

¹⁰³ RESTATEMENT (SECOND) OF TORTS § 552 (1986).

¹⁰⁴ Joel Rothstein Wolfson, *Electronic Mass Information Providers and Section 552 of the Restatement (Second) of Torts: The First Amendment Casts A Long Shadow*, 29 RUTGERS L.J. 67, 70 (1997) (“[The Second Restatement’s] Section 552 is not an accurate statement of the law of negligent misrepresentation for mass information providers, and mass distributors of information have almost never been held liable for inaccurate information, even where such information was justifiably and foreseeably relied upon in economic transactions. Most often, the courts simply refuse to impose liability under Section 552.”).

¹⁰⁵ An illustration to the Second Restatement provides: “The A Newspaper negligently publishes in one of its columns a statement that a certain proprietary drug is a sure cure for dandruff. B, who is plagued with dandruff, reads the statement and in reliance upon it purchases a quantity of the drug. It proves to be worthless as a dandruff cure and B suffers pecuniary loss. The A Newspaper is not liable to B.” RESTATEMENT (SECOND) OF TORTS § 552 cmt. a.

¹⁰⁶ See generally Lyrrisa Lidsky, *Cheap Speech and the Gordian Knot of Defamation Reform*, 3 J. FREE SPEECH L. 79 (2023).

potential physical harm or property damage caused by the vehicles' AI.¹⁰⁷ LLM users, in contrast, have much higher capacity (and a corresponding obligation, as we argue here) to avoid the potential reputational harm to others posed by their LLM usage.

3. Presumed damages

An alternate basis for the judge's dismissal in *Walters* was the fact that there was no evidence that ChatGPT's hallucination caused harm to Walters' reputation.¹⁰⁸ This holding highlights a curious feature of defamation law: the doctrine of "presumed damages," which allows factfinders to simply "presume" injury to plaintiffs in some types of defamation cases, based on the fact of a defamatory statement's publication.¹⁰⁹ This doctrine often permits plaintiffs to recover substantial damages without alleging or proving any actual harm to reputation or even any

¹⁰⁷ See Lior, *supra* note 89, at 95–96 (arguing for strict liability for AI-generated harms).

¹⁰⁸ Benjamin C. Zipursky, *Defamation, Presumed Damages, and Reputational Injury: A Legal and Philosophical Inquiry*, 4 J. FREE SPEECH L. 797, 831 (2024) (contending that the availability of presumed damages in defamation is justified because reputational injury involves the diminishment of ideas and attitudes that third parties hold about a person and such injury may be both instrumentally and intrinsically harmful). See also Order, *supra* note 46, at 19.

¹⁰⁹ See, e.g., *Palmtag v. Republican Party of Neb.*, 999 N.W.2d 573, 597 (Neb. 2024) (holding that plaintiff, a public figure, presented a prima facie case of libel per se and thus did not need to establish special damages because "[t]he common law of defamation allows recovery of purportedly compensatory damages without any evidence whatsoever of actual loss, if the defamation is deemed per se"); see also Steven A. Krieger, *Defamation Per Se Cases Should Include Guaranteed Minimum Presumed Damage Awards to Private Plaintiffs*, 58 SAN DIEGO L. REV. 641, 662–63 n.133 (2021) (establishing that forty states allow for some form of presumed damages for defamation per se in some types of cases).

emotional distress.¹¹⁰ The rationale for the doctrine is that injury to reputation occurs through subtle means not susceptible to easy proof.¹¹¹ How does one prove, for example, why one is no longer invited to friends' parties, why one's colleagues seem less friendly, or why one's job application was rejected? Allowing recovery of presumed damages is a way to relieve plaintiffs of onerous proof requirements in cases involving intangible but real harms. Yet common-law courts often view presumed

¹¹⁰ DAN B. DOBBS & CAPRICE L. ROBERTS, LAW OF REMEDIES § 7.2.2 (3d ed. 2018) ("At common law, courts imposed a kind of strict liability for libelous publications. Once defamatory content was shown, courts *presumed* that the defendant was at fault They also presumed that plaintiff had suffered damages. Plaintiffs were allowed to recover large sums as general damages without proof of either economic loss or any actual mental distress."). Juries are often instructed to award presumed damages based on no criteria other than the character of a defendant's statement and the circumstances of publication. *See, e.g.*, *Hennis v. O'Connor*, 388 N.W.2d 470, 476 (Neb. 1986) ("A proper instruction would have informed the jurors that if they found that the alleged defamatory statements had been made and that they were intended to identify the plaintiff as the thief, then the jury should determine the resulting damages."). Some jurisdictions do, however, instruct juries that their award must be reasonable. *See, e.g.*, *Arno v. Stewart*, 245 Cal. App. 2d 955, 969 (1966) ("The court then advised the jurors that the giving of instructions on damages should not be taken as an indication that liability was or was not established; and that damages should be reasonable. The instructions concluded with directions concerning the manner of deliberating and the use of proposed verdicts."); *see also* *Hoffmann v. Clark*, 975 N.W.2d 656, 668 (Iowa 2022) (providing more guidance to ensure a reasonable award and instructing jury to consider: "(1) What was the prior reputation of the plaintiff? (2) Did the plaintiff suffer emotional distress? (3) What type of defamatory statements were made? (4) How many defamatory statements were made? (5) How widely were they disseminated? (6) Over what period of time were they made and disseminated? (7) Were they ever retracted? (8) Was there evidence of bad faith?"). Presumed damages are not unique to defamation law. As Thomas Cotter points out, tort law typically presumes damages in cases involving trespass to real property as well, though it may limit these damages to nominal amounts. Thomas F. Cotter, *Standing, Nominal Damages, and Nominal Damages "Workarounds" in Intellectual Property After TransUnion*, 56 U.C. DAVIS L. REV. 1085, 1117 (2023).

¹¹¹ *See* DOBBS & ROBERTS, *supra* note 110, § 7.2.2. *See also* *Bongiovi v. Sullivan*, 138 P.3d 433, 450 (Nev. 2006) (upholding jury instructions stating that "[d]amages are presumed in slander per se actions because of the impossibility of affixing an exact monetary amount for present and future injury to the plaintiff's reputation, wounded feelings and humiliation, loss of business and any consequential physical illness or pain A slanderous communication constitutes slander per se if it would tend to injure the plaintiff in his trade, business, profession or office," and "[i]f you find defendant['s] remarks defamatory and if you further find these statements injured plaintiff's business or profession, then damages are presumed").

damages with suspicion.¹¹² Because presumed-damages awards are untethered to tangible criteria, they can end up being roulette-like windfalls far removed from actual loss.

The *Walters* judge conceded that Georgia law allows defamation plaintiffs who are falsely accused of serious crimes to rely on the doctrine of presumed damages, at least as an initial matter.¹¹³ The doctrine's presumption of damages, however, is rebuttable. Here, that presumption was rebutted by "undisputed facts."¹¹⁴ Walters did not even claim that he had suffered harm, and he admitted under oath that he did not.¹¹⁵ Moreover, Riehl, "the only person who received the challenged ChatGPT output," admitted that he was "always skeptical" of it and quickly established that it was untrue.¹¹⁶

This aspect of the *Walters* decision has important implications for future cases. OpenAI produced, during this litigation, a record of outputs to Riehl's prompts about Mark Walters. If OpenAI could likewise produce all the outputs ChatGPT had created containing the name Mark Walters, it would be a simple matter to establish the scope of harm to Walters' reputation. Here, OpenAI's record helped show that Riehl had no reason to believe the hallucination was true; in the future, plaintiffs could subpoena search records more comprehensively to show how many users or prompters received defamatory hallucinations about them.

4. The path not taken in *Walters*: Is ChatGPT a publisher or a distributor?

While the judge's order in *Walters* was impressively innovative in adapting defamation law to the circumstances, the order failed to address an important issue that is likely to arise in subsequent defamation-by-hallucination cases, namely

¹¹² Courts often restrict their recovery to the categories of cases deemed most likely to cause real harm. For example, courts typically require plaintiffs suing for slander to prove special damages (concrete out-of-pocket losses) unless the defendant's defamatory communication falls into one of four "per se" categories thought especially likely to cause reputational harm. In Georgia law, applied in the *Walters* case, a plaintiff need not prove special damages if the defendant's defamatory statement involves the "commission of a crime to another." Order, *supra* note 46, at 19 (citing *Riddle v. Golden Isles Broad., LLC*, 292 Ga. App. 888, 891 (2008)).

¹¹³ *Id.*

¹¹⁴ *Id.*

¹¹⁵ *Id.* at 18.

¹¹⁶ *Id.* at 21.

whether OpenAI is a “publisher” of the outputs ChatGPT provides in response to user queries. “[P]ublication is a term of art”¹¹⁷ in defamation law. Because reputation is the esteem in which one is held by others, reputational harm stems from the effect of defamatory statements on, well, others. Thus, at least one person other than the plaintiff must receive the defendant’s defamatory communication for it to be actionable,¹¹⁸ and the defendant must typically communicate the defamatory matter intentionally or negligently.¹¹⁹ Though it is impossible for a generative AI model to have even general intent, it seems likely that courts might treat the underlying model design as evincing OpenAI’s intent to communicate. But that is not the end of the story, because the element of publication has an additional dimension that is rooted in public policy concerns.

Publication is treated differently depending on whether the person making the defamatory statement is a “publisher”¹²⁰ or a “distributor.”¹²¹ Publishers are those who write, develop, select, prepare, and edit original content.¹²² The category includes individual speakers, newspapers, people posting on social media, and book authors and editors.¹²³ The law imputes to this class of defamation defendants

¹¹⁷ *Ostrowe v. Lee*, 175 N.E. 505, 505 (N.Y. 1931) (citing treatises on defamation for the proposition).

¹¹⁸ See RESTATEMENT (SECOND) OF TORTS § 577(1) (1986); Draft Restatement, *supra* note 32, § 2 cmt. c.

¹¹⁹ Draft Restatement, *supra* note 32, § 2.

¹²⁰ See *id.* § 4 cmt. b. Publishers are sometimes referred to as “primary publishers,” as opposed to distributors, which are referred to as “secondary publishers.” See Amanda Groover Hyland, *The Taming of the Internet: A New Approach to Third-Party Internet Defamation*, 31 HASTINGS COMM’NS & ENT. L.J. 79, 96–97 (2008) (“Primary publishers that republish false statements are usually held to the same standard of liability as the original author of the statement.”).

¹²¹ See generally Draft Restatement, *supra* note 32, § 4. See also, e.g., *Barrett v. Rosenthal*, 146 P.3d 510, 519 (Cal. 2006) (asserting that the common law imposes different standards on “‘primary publishers,’ who have control over content, and ‘distributors,’ who do not”).

¹²² See generally Draft Restatement, *supra* note 32, § 4 cmt. b; see also RESTATEMENT OF THE LAW OF TORTS § 581 (1934) (stating that one who disseminates defamatory material originally published by a third person is liable “unless he has no reason to know of its defamatory character”); RESTATEMENT (SECOND) OF TORTS § 581(1) (1986) (stating that “one who only delivers or transmits defamatory matter published by a third person is subject to liability if, but only if, he knows or has reason to know of its defamatory character”).

¹²³ See Draft Restatement, *supra* note 32, § 4 cmt. b.

knowledge of the contents they publish and imposes on them a duty to take reasonable steps to confirm the truth of what they publish—even if they are repeating verbatim the words of others.¹²⁴ The basis of the duty is their active role in content creation, their ability to control the content of what they publish before they publish it, their knowledge of the need to do so, and the foreseeability of harm if they do not.¹²⁵

At first glance, OpenAI looks like a publisher. After all, the producers of LLMs “use software to *create* messages rather than merely hosting them.”¹²⁶ However, the producers of LLMs find it “difficult or impossible to know in advance what output an LLM is likely to generate, and similarly difficult or impossible to entirely anticipate undesirable outputs.”¹²⁷ Moreover, as legal scholar Jane Bambauer perceptively argues, LLM producers differ from traditional publishers in that “the crafting of outgoing messages is done on a large scale and based on automated procedures

¹²⁴ *Id.*

¹²⁵ Publishers “can be held liable for defamatory statements contained in their works even absent proof that they had specific knowledge of the statement’s inclusion.” *Zeran v. Am. Online, Inc.*, 129 F.3d 327, 331 (4th Cir. 1997) (citing *W. PAGE KEETON ET AL., PROSSER AND KEETON ON TORTS* § 113 (5th ed. 1984)). *See also* *Smith v. Utley*, 65 N.W. 744, 745 (Wis. 1896) (holding that “liability attaches to the editor [of a defamatory statement] under the theory that the matter is constructively under his supervision,” and thus disregarding editor’s testimony that he did not “authorize or know of” the defamatory communication); *Overstock, Inc. v. Gradient Analytics, Inc.*, 61 Cal. Rptr. 3d 29, 48 (Ct. App. 2007) (“One who takes a responsible part in a publication of defamatory material may be held liable for the publication.”). *Cf. Sakamu v. Zellerbach Paper Co.*, 77 P.2d 313, 320 (Cal. Ct. App. 1938) (holding that a newspaper’s business manager was not a publisher because he lacked “any control over the editorial staff” and was instead “directly concerned in the *distribution* of the newspaper” (emphasis added); *see also id.* (“It is a good defense for the vendor or distributor of a newspaper or periodical to show that he had no knowledge of the libelous matter, and that there were no extraneous facts which should have put him on his guard.” (citing 37 C.J.S. *Libel and Slander* § 307 (1925))).

¹²⁶ Bambauer, *supra* note 45, at 352 (emphasis in original).

¹²⁷ Burk, *supra* note 37, at 194 (“As a result of the system’s development of its own algorithm, it is also often difficult or impossible to determine exactly how the system reaches a particular outcome—the inner workings of the system are complex and independently determined, creating an obscured or impenetrable ‘black box.’ Consequently, in the case of LLM output, it may be difficult or impossible to know in advance what output an LLM is likely to generate, and similarly difficult or impossible to entirely anticipate undesirable outputs.”).

set in advance.”¹²⁸ To extend Bambauer’s analysis, creating an AI chatbot is more akin to creating an Internet search engine than it is to producing content for a newspaper.¹²⁹ In both instances, the information producers typically lack both advance knowledge of the need to control a particular piece of content and the practical ability to implement such control.¹³⁰ Unlike a newspaper editor, ChatGPT neither independently decides what topics to cover nor exercises control over discrete content; instead, it responds to prompts by collating content as its algorithms dictate.¹³¹ It is a tool for producing content rather than a content producer.¹³²

Arguably, therefore, ChatGPT is more like a “distributor” or other “innocent disseminator” of information than a publisher.¹³³ The common-law distributor category includes libraries, bookstores, and newsstands; the law also relieves telegraph operators and printers of liability for “innocent dissemination.”¹³⁴ Distributors are

¹²⁸ Bambauer, *supra* note 45, at 352.

¹²⁹ *Id.* at 355. As Reuben Binns and Lilian Edwards explain: “Search engines can be seen as ‘small language models,’ predecessors to LLMs, with queries analogous to prompts and search results to outputs.” Reuben Binns & Lilian Edwards, *Reputation Management in the ChatGPT Era* 5 (Sep. 12, 2024), <https://perma.cc/TLA2-GFWY>.

¹³⁰ The EU’s AI Act describes AI developers as “providers.” Under the EU’s terminology, “users” are natural or legal persons that deploy an AI system in a professional capacity, and “end-users” are the persons (presumably) prompting the LLM. Council Regulation 2024/1689, 2024 O.J. (L 1689).

¹³¹ At least in its early iterations, ChatGPT had little ability to verify the “facts” its algorithms produced, although chain of thought and chain of verification reasoning models are changing this, as we explain in the next section.

¹³² As Reuben Binns and Lilian Edwards point out, however, one should not succumb to “the ‘automation fallacy,’ prevalent in early Internet cases, that algorithms do their own thing in some unforeseeable way, rather than as a result of deliberate programming and fine-tuning by their owners with very careful (and commercially lucrative) optimization goals.” Binns & Edwards, *supra* note 129, at 5.

¹³³ Reuben Binns and Lilian Edwards survey English, Australian, Canadian, and German precedent regarding search engines as publishers to show that the answer to whether OpenAI is a publisher is not clear: “Although the precedents to date are mixed, it seems hard to justify treating automated outputs as speech with constitutional value on the one hand, and deny the possibility for damaging reputations on the other.” *Id.* at 4.

¹³⁴ *Layton v. Harris*, 3 Del. 406, 407 (Super. Ct. 1842) (holding a post-master’s “innocent delivery of a sealed letter” is not a publication unless the post-master knows of the letter’s contents);

those who merely transmit, deliver, or make available communicative content that has been produced by others.¹³⁵ They do not actively create or collate this content.

Misut v. Mooney, 475 N.Y.S.2d 233, 235–36 (Sup. Ct. 1984) (holding that it “does not view duty of a printer to be inclusive of an obligation to confirm facts, check sources and to thereby be responsible for the truth of printed statements. To do so would establish the printer as a censor. It would be impractical in economic terms and undesirable in social terms. In short, it would be bad policy.”); Maynard v. Port Publ’ns, Inc., 297 N.W.2d 500, 507 (Wis. 1980) (stating that imposing liability on distributors, such as a “quick and inexpensive printing service,” for “failure to inspect” would “of necessity” make them “censors,” which would both make their services more expensive and discourage them from making controversial information available, causing “a deleterious effect on the free dissemination of information which is fundamental in our society”); *but see* Holden v. Am. News Co., 52 F. Supp. 24, 35 (E.D. Wash. 1943) (asserting that “[w]hether lack of knowledge [on the part of a distributor] is a defense is doubtful”). It is important to note that distributor cases in American law are rare. *See* Draft Restatement, *supra* note 32, § 4 & Reporter’s Notes (including an explanation of the evolution of the common law’s innocent dissemination defense into the modern liability rules governing distributor liability).

¹³⁵ *Id.* § 4 cmt. b & Reporter’s Notes (explaining rationale for distributor rules in comment b, and collecting “innocent dissemination” distributor cases and explaining the evolution of distributor rules in Reporter’s Notes to § 4). For distributor and innocent disseminations, see Carafano v. Metrosplash, Inc., 207 F. Supp. 2d 1055, 1073 (C.D. Cal. 2002) (“A distributor of defamatory matter is blameless if the distributor has no notice of its possible falsity.”); Bowerman v. Detroit Free Press, 283 N.W. 642, 645 (Mich. 1939) (“It is a good defense for a vendor or distributor of a newspaper or periodical to show that he had no knowledge of the libelous matter and that there were no extraneous facts which should have put him on his guard.”); Balabanoff v. Fossani, 192 Misc. 615 (N.Y. Sup. Ct. 1948) (denying plaintiff’s motion to strike a news vendor’s innocent dissemination defense); Albi v. St. & Smith Publ’ns, 140 F.2d 310, 313 (9th Cir. 1944) (stating, in a case involving a magazine distributor, “Every person who sells or gives away a written or printed copy of a libel may be made a defendant, unless, indeed, he can satisfy the jury that he was ignorant of the contents. The onus of proving this lies on the defendant.”); Lerman v. Chuckleberry Publ’g, Inc., 521 F. Supp. 228, 235 (S.D.N.Y. 1981) (“With respect to distributors, the New York courts have long held that vendors and distributors of defamatory publications are not liable if they neither know nor have reason to know of the defamation.”), *rev’d on other grounds sub nom.* Lewis v. Time Inc., 83 F.R.D. 455, 463 (E.D. Cal. 1979) (stating that there is “no California case imposing liability where a distributor merely sold an unchanged libelous periodical” and that the “common thread” of extant case law is that “there can be no liability absent scienter”), *aff’d*, 710 F.2d 549 (9th Cir. 1983); Hartmann v. Am. News Co., 69 F. Supp. 736, 739 (W.D. Wis. 1947) (sending to the jury the issue of whether a distributor was negligent in failing to know of the defamatory statements), *aff’d*, 171 F.2d 581, 585 (7th Cir. 1948); Spence v. Flynt, 647 F. Supp. 1266, 1273 (D. Wyo. 1986) (allowing defamation action against a distributor to proceed because of “very special and rare circumstances” regarding distributor’s knowledge of libelous statements); Dworkin v. L.F.P., 647 F. Supp. 1275, 1277–78 (D. Wyo.

Instead, they relay to others content over which they have little control in advance; in fact, many distributors have little ability to edit content, such as a library book, after being notified that the content is defamatory; they can only block future access to the disputed portion or remove the whole work. Moreover, whereas publishers implicitly warrant the truth of what they publish, distributors, of necessity, do not.¹³⁶ The law does not impose a duty on the distributor to verify or edit content prior to its “communication,” because it is economically impractical and socially undesirable that the distributor do so.¹³⁷

At their core, publisher and distributor rules are rooted in communications policy concerns, and ChatGPT is not the first communications tool to resist easy categorization. Although ChatGPT’s “generative” capability differentiates it from other distributors, the fact that it creates content without prior human editorial control makes it a poor fit for the publisher category.

In categorizing ChatGPT as either publisher or distributor, a key policy concern should be the need to protect the transformative potential of AI models as communications tools. As we will see in the next section, AI model hallucinations are inevitable: Hallucinations arise from the probabilistic nature of AI “reasoning,” incomplete or conflicting AI training data, lack of human understanding, and misinterpretation of users’ prompts.¹³⁸ Making developers responsible for every defamatory hallucination could result in crippling liability, even for the wealthiest AI

1986) (holding that distributor had not been fraudulently joined in plaintiff’s defamation action: “This was simply not a case of an innocent magazine seller unwittingly disseminating allegedly libelous material. . . . These facts do not establish that Park Place Market lacked scienter for libel.”); *Macaluso v. Mondadori Publ’g Co.*, 527 F. Supp. 1017, 1019 (E.D.N.Y. 1981) (allowing libel claim against distributor to proceed to trial for jury determination of innocent dissemination issue).

¹³⁶ See, e.g., *Osmond v. EWAP, Inc.*, 200 Cal. Rptr. 674, 680 (Ct. App. 1984) (holding that an adult bookstore was subject to the rule governing distributors: “One who merely plays a secondary role in disseminating information published by another, as in the case of libraries, news vendors, or carriers, may avoid liability by showing there was no reason to believe it to be a libel. It is, therefore, a good defense for a mere vendor or distributor of a newspaper or other periodical to show that he had no knowledge of the libelous matter and that there were no extraneous facts which should have put him on guard.”).

¹³⁷ *Misut*, 475 N.Y.S.2d at 235–36 (stating that printers have no duty to confirm the truth of what they publish); *Maynard*, 297 N.W.2d at 507 (stating that imposing liability on distributors for “failure to inspect” would have a negative effect on the free flow of information).

¹³⁸ See discussion *infra* in Part II.B.

producers. While only a few libel actions have been filed thus far, every successful suit has the potential to inspire others, and whether the proverbial dam breaks is partially dependent on the rules laid down in these early cases. Moreover, not all AI developers are multibillion-dollar entities, and getting the liability rules “wrong” has the potential not only to block the entry of smaller AI developers into the industry but also to disincentivize innovations that would improve accuracy in the long run, as we show in the next section.

Treating AI models as publishers also risks encouraging some users to try to generate defamatory hallucinations in order to pursue litigation against AI producers. The prompter in the *Walters* case kept prompting until he produced a defamatory output, even in the face of red flags warning that the output would be unreliable. Walters, as plaintiff, then sued the defendant with the deep pockets, not the prompter. Legal rules should not incentivize prompters to go looking for defamation, but instead should encourage them to behave reasonably.

Tort law often seeks to place liability on the “cheapest cost avoider”—the party who can prevent harms to others at the lowest cost.¹³⁹ Placing liability on the cheapest cost avoider promotes efficient resource allocation.¹⁴⁰ In the *Walters* case, the cheapest cost avoider was the prompter who had the actual facts at his fingertips but ignored them. Indeed, one might view that prompter as playing such an active role that he was an independent, intervening cause in the publication process. The prompter, a journalist, relied on an unreliable source without verification and could have easily avoided any potential reputational harm to Walters simply by verifying

¹³⁹ Guido Calabresi, *Views and Overviews*, 1967 U. ILL. L.F. 600, 608 (introducing the concept of the cheapest cost avoider); GUIDO CALABRESI, *THE COSTS OF ACCIDENTS: A LEGAL AND ECONOMIC ANALYSIS* 155 (1970) (“[T]he search for the cheapest avoider of accident costs is the search for that activity which has most readily available a substitute activity that is substantially safer. It is a search for that degree of alteration or reduction in activities which will bring about primary accident cost reduction most cheaply.”); Guido Calabresi & Spencer Smith, *On Tort Law’s Dualisms*, 135 HARV. L. REV. F. 184, 185 (2022) (“What constitutes a civil wrong most often derives from the regulatory needs of society, and hence often from a desire to place liability on the “cheapest cost avoider.” But what is “cheap” and what is “costly” itself derives from the tastes and values of society, which can be influenced by the current set of civil wrongs.”).

¹⁴⁰ See *id.* See also Catherine M. Sharkey, *Modern Torts: Preventing Harms, Not Recognizing Wrongs*, 134 HARV. L. REV. 1423, 1425 (2021) (discussing the primacy of cheapest-cost-avoider analysis in tort law).

ChatGPT's output. Had Walters actually suffered harm, Riehl was an active contributor to that harm.

Of course, not every case will involve a prompter goading a chatbot until it produces errors and persisting in the face of warnings about such errors. A recent Minnesota case presents a much more challenging scenario, because it did not involve unreasonable prompting behavior but instead involved unreasonable reliance on a false output.¹⁴¹ In *LTL LED, LLC v. Google LLC*, Google's AI Overview, which is part of its Search feature, falsely stated that the Minnesota Attorney General had sued a solar panel company for deceptive sales practices.¹⁴² Google's AI Overview provided four sources supposedly leading to its "summary" but, according to the complaint, none of them supported the AI's assertion about the deceptive-sales lawsuit—which did not exist.¹⁴³ *LTL* is a far more compelling case than *Walters*, because *LTL* lists in its complaint specific examples of contracts and prospective business relationships that it lost because of the defamatory allegations.¹⁴⁴ If the allegations in the complaint are true, *LTL* may represent a case in which prompters did not (re)publish unverified information but nonetheless acted upon it, causing provable harm to the plaintiff.

Given the complicated dynamics of these cases, what liability rule should attach to ChatGPT outputs? To answer that question, it is helpful to delve more deeply into the history of defamation's distributor rules. The Second Restatement of Torts asserts, and many cases have repeated, that distributors will be liable for third-party content only if they "know[] or ha[ve] reason to know of its defamatory character."¹⁴⁵ Distributor case law, however, is "underdeveloped,"¹⁴⁶ and cases, many of

¹⁴¹ Notice of Removal exh. 1 [Complaint], *LTL LED, LLC v. Google LLC*, No. 0:25-cv-02394 (D. Minn. June 9, 2025).

¹⁴² *Id.* at 2.

¹⁴³ *Id.* at 3.

¹⁴⁴ *Id.* at 19–20.

¹⁴⁵ RESTATEMENT (SECOND) OF TORTS § 581(1) (1986) ("[O]ne who only delivers or transmits defamatory matter published by a third person is subject to liability if, but only if, he knows or has reason to know of its defamatory character.").

¹⁴⁶ Benjamin C. Zipursky, *Online Defamation, Legal Concepts, and the Good Samaritan*, 51 VALPARAISO U. L. REV. 1, 21–22 (2016) (observing that the formal rule of distributor liability is "underdeveloped" and chiefly dependent upon case law developed prior to the Restatement of Torts).

which predate the Second Restatement, tend to repeat the English rule that a distributor may become liable as a publisher if the distributor has, or should have had, knowledge of the content of the defamatory communication they transmitted.¹⁴⁷ This rule grew out of English cases that created a privilege of “innocent dissemination” for distributors such as bookstores.¹⁴⁸ The Second Restatement rule¹⁴⁹ simply switched the burden of proving culpable dissemination to the plaintiff.¹⁵⁰

A careful reading of the cases on distributors suggests that the actual rule applied in distributor cases does not necessarily condemn distributors for failure to take down defamatory content upon notice but instead recognizes that liability should not flow where takedown is not feasible. Neither a bookstore nor library has ever been held liable for defamation in the history of U.S. common law.¹⁵¹ This outcome supports society’s interest in safeguarding public access to information because the only recourse such distributors have, when notified that a work in their collection contains defamatory content, is to remove the entire work rather than

¹⁴⁷ See, e.g., *Bowerman v. Detroit Free Press*, 283 N.W. 642, 645 (Mich. 1939) (repeating the rule).

¹⁴⁸ For U.S. cases applying the innocent dissemination defense, see *Balabanoff v. Fossani*, 192 Misc. 615 (N.Y. Sup. Ct. 1948); *Albi v. St. & Smith Publ’ns, Inc.*, 140 F.2d 310, 313 (9th Cir. 1944). For an explanation of the origins of the innocent dissemination defense, see generally Rosemary Tobin, *Publication and Innocent Dissemination in the Law of Defamation: Adapting to the Internet Age*, 27 N.Z.U. L. REV. 102 (2016). The origin of the doctrine is the English case *Emmens v. Pottle* [1884] 16 QBD 354. *Emmens* involved newspapers vendors. Lord Esher described them as “innocent disseminators of a thing which they were not bound to know was likely to contain a libel.” *Id.* at 357. In *Emmens*, it was stated that newspaper vendors could avoid liability by showing they neither knew nor should have known of the defamatory contents in their newspapers. *Id.* at 358. The *Emmens* doctrine—which developed to mitigate the harshness of strict liability—has been widely influential. One of the few cases imposing liability under *Emmens*’ “innocent dissemination” rule is *Vizetelly v. Mudie’s Select Libr., Ltd.* [1900] 2 QB 170. In upholding a jury verdict for the plaintiff, the judges in *Vizetelly* distinguished *Emmens* on the ground that the vendor there had no practical means of becoming aware of defamatory content prior to publication. See *Oriental Press Grp. Ltd. v. Feva-works Sols. Ltd.* [2013] 16 H.K.C.F.A.R. 366, ¶ 24 (explaining, in a case involving an internet service provider, that the doctrine arose in part because “[t]he strictness of the publication rule plainly called for some relaxation”).

¹⁴⁹ See, e.g., *Carafano v. Metrosplash, Inc.*, 207 F. Supp. 2d 1055, 1073 (C.D. Cal. 2002) (“A distributor of defamatory matter is blameless if the distributor has no notice of its possible falsity.”).

¹⁵⁰ See distributor cases cited *supra*, note 135.

¹⁵¹ See Draft Restatement, *supra* note 32, § 4 cmt. c (summarizing case law research).

selectively editing the defamatory portion. Based on reported U.S. cases, it appears that a newspaper or magazine vendor has only once, in 1943, been held liable for distributing defamatory content, though several distributors have faced trial over whether they were “ignorant of the contents”¹⁵² of the allegedly defamatory matter they distributed. A survey of these cases indicates that courts have applied the distributor rule to protect the role distributors play in providing broad public access to information created by others. That role would be compromised if distributors were forced to remove content simply upon notice.

A survey of this complex case law led the co-reporters of the Third Restatement of Defamation, which was approved by the ALI Council in October 2025, to rearticulate the distributor rule as follows: Section 4(b) of the draft Restatement’s distributor rule provides that a distributor is only deemed responsible for publishing a defamatory communication if “the distributor (1) has reason to know that the communication contains defamatory content; and (2) has authority to edit the communication or cease distributing it and fails to do so under circumstances that would give rise to a reasonable inference that the distributor endorses or ratifies the defamatory content.”¹⁵³ Applied thoughtfully, this rule has significant implications for generative AI developers.

AI developers have strong incentives to reduce hallucinations, and not just because defamation lawsuits are time-consuming and expensive. Every defamation case casts doubt on the reliability of AI outputs.¹⁵⁴ This is doubtless why Meta not only settled with defamation-by-hallucination victim Robby Starbuck but also is-

¹⁵² See, e.g., *Albi*, 140 F.2d at 313 (refusing to dismiss magazine distributor from defamation suit because defendant bore the burden of proving innocent dissemination).

¹⁵³ See Draft Restatement, *supra* note 32, § 4 cmt. c.

¹⁵⁴ This article assumes that holding ChatGPT and other AI producers liable for presumed damages for every defamatory hallucination their LLMs produce runs the risk of stifling innovation. An excellent counterresponse is that only a few suits have been filed thus far, even though LLMs are currently being used by millions of people; moreover, AI companies are multi-million- or -billion-dollar businesses capable of internalizing the costs they impose on defamation victims. However, just because few suits have been filed so far does not mean few suits will be filed in the future, since every reported lawsuit potentially inspires others. Thus, it is important for the law to get these early cases “right.” While many AI companies, at present, are capable of paying large damage awards, such awards should always be justified, and smaller startups may be deterred from entering the playing field by the eventual volume of claims and their potential magnitude.

sued an apology for the hallucination and promised to work with Starbuck to eliminate such mistakes in the future.¹⁵⁵ As detailed in the next section, new reasoning models incorporating content verification in their design represent an attempt by model developers to solve the hallucination problem. Even so, AI developers cannot prevent hallucinations entirely, and the reasons these errors occur in any individual instance are not always clear. In fact, a prompter cannot necessarily replicate a hallucination with the same prompt on a subsequent occasion. Developers refer to the opacity of the mechanism by which AI hallucinations occur as the “black box.”¹⁵⁶ The “black box” effect can make blocking challenging or even impossible for AI developers.¹⁵⁷ Although in some instances ChatGPT may be able to check (and block) generated output using words known to be associated with a specific defamatory assertion, in other instances the only way to completely prevent such hallucinations will be to disable all searches for a particular topic.¹⁵⁸ And disabling searches will result in reduced public access to information and encourage strategic takedown requests, especially by the wealthy and powerful.¹⁵⁹

One might counter that AI developers should not be treated similarly to distributors because a generative AI model is not a truly “passive” conduit in the same sense as a library, bookstore, or newspaper. AI developers, after all, create algorithms that “generate” content, though the developer is not active at the moment

¹⁵⁵ Amalia Huot-Marchand, *Meta Settles with Conservative Activist over AI Chatbot Lawsuit*, HILL (Aug. 8, 2025), <https://perma.cc/8ZTH-JW8D>. Subsequently, Starbuck sued Google alleging its AI platforms produce similar defamatory hallucinations about him that had been viewed by almost 3 million users. See Taylor Penley, *Google Hit with Lawsuit over AI ‘Hallucinations’ Linking Conservative Activist to Child Abuse Claims*, FOX NEWS (Oct. 28, 2025, 8:10 am), <https://www.foxnews.com/media/google-hit-lawsuit-over-ai-hallucinations-linking-conservative-activist-child-abuse-claims>.

¹⁵⁶ See generally Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 889, 892 (2018) (“[I]t may not be possible to truly understand how a trained AI program is arriving at its decisions or predictions.”).

¹⁵⁷ “See *supra* note 155.

¹⁵⁸ See, e.g., Jibin Joseph, *Apple Pauses AI Notification Summaries that Spread Fake News*, PCMag (Jan. 17, 2025), <https://perma.cc/ECJ3-X943>.

¹⁵⁹ For further elaboration of how notice-and-takedown regimes can lead to strategic behavior by the wealthy and powerful, see generally Jeffery Cobia, *The Digital Millenium Copyright Act Takedown Notice Procedure: Misuses, Abuses, and Shortcomings of the Process*, 10 MINN. J.L. SCI. & TECH. 387, 390 (2009) (citations omitted).

of generation. In response to the argument that AI models are not passive conduits, it is worth considering the common law's creativity in adapting to the development of the telegraph. A telegraph operator has actual knowledge of the content being transmitted, as well as the ability to decide whether to relay that content. A telegraph operator, in some sense, originates the content transmitted by translating it into the requisite code for transmission, and the recipient operator must decode it before delivering it to its intended recipient. Thus, a telegraph is not a passive conduit in the same sense a bookstore is.

But had the common law imposed publisher liability on telegraph operators, the development and deployment of this revolutionary technology would have been stymied. Telegraphs were the first medium, aside from the vastly more unreliable heliograph,¹⁶⁰ to enable almost instantaneous communication over long distances. Rather than allow defamation law to squander the telegraph's potential, common-law courts cleverly adapted existing legal rules to insulate telegraph companies from liability,¹⁶¹ as detailed in an excellent 1920 article by scholar and later dean of Columbia Law School Young B. Smith.¹⁶²

The courts did so by recognizing an "immunity" or privilege from liability for telegraph companies unless the telegraph operator "happened to know that the message was spurious or that the sender was acting, not in the protection of any legitimate interest, but in bad faith and for the purpose of traducing another."¹⁶³

¹⁶⁰ Joseph H. Sommer, *Against Cyberlaw*, 15 BERKELEY TECH. L.J. 1145, 1162 n.48 (2000) ("Note that the technological basis for heliography dates back to Archimedes, who focused the sun on enemy ships, the better to burn them. The limiting factor was social—the human and institutional infrastructure for a network of signaling stations.").

¹⁶¹ Telegraph companies were essential distributors of communications with actual knowledge of the contents they distributed. To protect telegraphs from crippling liability for "republishing" of defamatory falsehoods, courts at the turn of the last century developed ad hoc privileges to protect telegraphs and discourage them from censorship. See, e.g., Young B. Smith, *Liability of a Telegraph Company for Transmitting a Defamatory Message*, 20 COLUM. L. REV. 369 (1920); *Peterson v. W. Union Tel. Co.*, 67 N.W. 646 (Minn. 1896); *Nye v. W. Union Tel. Co.*, 104 F. 628 (C.C.D. Minn. 1900); *W. Union Tel. Co. v. Cashman*, 149 F. 367 (5th Cir. 1906); *Grisham v. W. Union Tel. Co.*, 142 S.W. 271 (Mo. 1911); *Klein v. W. Union Tel. Co.*, 13 N.Y.S.2d 441 (App. Div. 1939); *Flynn v. Reinke*, 225 N.W. 742, 743 (Wis. 1929); *W. Union Tel. Co. v. Brown*, 294 F. 167 (8th Cir. 1923).

¹⁶² Smith, *supra* note 161.

¹⁶³ *O'Brien v. W. Union Tel. Co.*, 113 F.2d 539, 543 (1st Cir. 1940). For a listing of early cases applying the telegraph operator privilege, see *Recent Cases*, 43 HARV. L. REV. 131, 144 (1929).

This immunity, or privilege, was “broad enough to enable the company to render its public service efficiently and with dispatch.”¹⁶⁴ Courts recognized the need of telegraph operators to swiftly convey a high volume of messages. Forcing telegraph operators to review every single message for defamatory content would be impractical and would cripple the operation of this important medium of communication.¹⁶⁵ Moreover, imposing liability on the telegraph company was unnecessary when the more culpable party was the sender.

The common law initially handled the telegraph problem as a question of privilege, not as a question of publication. Regardless, the point holds: Courts sometimes adapt common law liability rules to absolve conduits that are not entirely “passive,” at least where courts deem it necessary to achieve communications policy goals. It is admittedly a stretch to treat AI developers as distributors and shield them from liability unless they receive notice of the defamatory content and then fail to act reasonably in blocking it. But it may be a desirable stretch, at least at this point in AI’s development.

II. A MOVING TARGET: OPENAI’S EVOLUTION FROM FOUNDATION MODELS TO REASONING

The version of ChatGPT prompted in the *Walters* complaint, GPT-3.5, was what is known as a foundation model. Foundation models are pre-trained on enormous, diverse datasets and are designed to respond to a wide array of disparate tasks with a high degree of accuracy.¹⁶⁶ Foundation models include OpenAI releases such

¹⁶⁴ *O’Brien*, 113 F.2d at 543.

¹⁶⁵ *Id.* at 542 (“If the telegraph companies are to handle such a volume of business expeditiously, it is obvious that their agents cannot spend much time pondering the contents of the messages with a view to determining whether they bear a defamatory meaning, and if so, whether the sender might nevertheless be privileged.”).

¹⁶⁶ Amazon, *What Are Foundation Models?*, AMAZON WEB SERVS., <https://perma.cc/923C-FMZ8>. See Dahl et al., *supra* note 24, at 65, 88. Compare the general function of foundation models with proprietary AI platforms trained to execute a narrow scope of tasks. For example, “PrivateAI” is a legally trained large language model that can be used with confidential information. The law firm White & Case uses PrivateAI for confidential matters and allows attorneys to use publicly available LLMs for tasks not involving confidential data. Becky Yerak, *AI Comes to Bankruptcy, as Courts Deal with Its Ethical and Privacy Risks*, WALL ST. J. (Apr. 15, 2024). See also *Professional Class AI*, HARVEY AI, <https://perma.cc/67U2-BVAJ>.

as GPT-3, GPT-3.5, GPT-4, and GPT-4o. These models act as a “base,” or foundation, for a range of applications, and users build “on top of” them to develop separate applications better-suited for more specific, narrow purposes.¹⁶⁷

OpenAI released ChatGPT, a chatbot interface built on its GPT foundation, on November 30, 2022. ChatGPT quickly became the “fastest-growing consumer application in history,”¹⁶⁸ reaching 123 million users in just under three months.¹⁶⁹ Since 2022, OpenAI has recorded 5.5 billion page visits,¹⁷⁰ and the development of new AI models, sometimes called “frontier” models,¹⁷¹ has dramatically increased the “reasoning” power available to users. Along with this increased reasoning power come even more regulatory challenges. In this Part, we document the rapid changes taking place and explain why both the pace of the changes and the changes themselves—particularly the development of “chain-of-thought” (CoT) reasoning models—amplify these challenges.

¹⁶⁷ Elliot Jones, *What Is a Foundation Model?*, ADA LOVELACE INST. (July 17, 2023), <https://perma.cc/9YLD-M26X>. An illustrative example may be DALL-E, OpenAI’s image-generation model. GPT-3 is a foundation model. DALL-E was built on top of GPT-3’s language model, for the specific purpose of generating images from text prompts. DALL-E is itself a “foundational model,” and can be fine-tuned for more specific applications. Any bias present in GPT-3 may be present in DALL-E, and so forth. See *Understanding the Architecture of DALL.E*, HOWDEV, <https://perma.cc/47VF-YLGL>.

¹⁶⁸ Ben Wodecki, *UBS: ChatGPT May Be the Fastest Growing App of All Time*, AI BUS. (Feb. 3, 2023), <https://perma.cc/888X-8LY3>; see also Catherine Thorbecke, *A Year After ChatGPT’s Release, the AI Revolution Is Just Beginning*, CNN (Nov. 20, 2023), <https://perma.cc/Z6SN-BLKV>.

¹⁶⁹ Wodecki, *supra* note 168.

¹⁷⁰ *ChatGPT, DeepSeek, Character.ai Lead Global GenAI Web Apps*, FINTECH NEWS SWITZ. (June 20, 2025), <https://perma.cc/G5AB-78WA>.

¹⁷¹ The term “frontier model” was coined by technologists to refer to “highly capable foundation models that could possess dangerous capabilities sufficient to pose severe risks to public safety.” Markus Anderljung et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety* (Nov. 7, 2023), <https://perma.cc/8UQL-QCPT>. For criticism of the terminology, see Gina Helfrich, *The Harms of Terminology: Why We Should Reject So-Called “Frontier AI”*, 4 AI & ETHICS 699 (2024).

A. A Chain-of-Thought Timeline

OpenAI released a preview of its “reasoning” model, o1, on September 12, 2024.¹⁷² Shortly thereafter, DeepSeek,¹⁷³ Google,¹⁷⁴ and Anthropic¹⁷⁵ released their own reasoning models. Reasoning models depart from foundation models in important respects. OpenAI’s product announcement presented o1 as a “significant advancement” and a “new level of AI capability.”¹⁷⁶ This new capability stems in part from increases in AI’s “context window” (the amount of text the LLM can process at one time),¹⁷⁷ attention (the mechanism by which the AI focuses selectively on different data inputs),¹⁷⁸ and output length.¹⁷⁹ According to OpenAI and

¹⁷² *Introducing OpenAI o1*, OPENAI, <https://perma.cc/2Z6D-4E2X>.

¹⁷³ DeepSeek released its open-source reasoning model, DeepSeek-R1, on January 22, 2025. DeepSeek claims its R1 model performs comparably to OpenAI-o1-1217 on reasoning tasks. DeepSeek-AI, *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning* (Jan. 22, 2025), <https://perma.cc/MC49-M6XU>.

¹⁷⁴ Google has released a series of reasoning models under its Gemini 2.0 model family. Its most powerful model (at the time of this writing) is Gemini 2.0 Pro. As of February 5, 2025, Google has made its first reasoning model, 2.0 Flash Thinking Experimental free-to-access. Koray Kavukcuoglu, *Gemini 2.0 Is Now Available to Everyone*, GOOGLE: KEYWORD (Feb. 5, 2025), <https://perma.cc/DWC5-8V6J>.

¹⁷⁵ On February 24, 2025, Anthropic released Claude 3.7 Sonnet, a “hybrid reasoning model.” As of March 1, 2025, the “extended thinking mode,” with visible thought blocks, is available only to paid subscribers. *Claude 3.7 Sonnet and Claude Code*, ANTHROPIC (Feb. 24, 2025), <https://perma.cc/UD9N-BBY9>.

¹⁷⁶ *Introducing OpenAI o1-Preview*, OPENAI (Sep. 12, 2024), <https://perma.cc/P8KP-FXHR>.

¹⁷⁷ Context window describes “the amount of text, in tokens, that the model can consider or ‘remember’ at any one time. A larger context window enables an AI model to process longer inputs and incorporate a greater amount of information into each output.” Dave Bergmann, *What Is a Context Window?*, IBM (Nov. 7, 2024), <https://perma.cc/QG4S-LJB5>.

¹⁷⁸ “Attention mechanisms allow the model to focus on keywords in a sentence, much like how humans focus on key words in conversations.” Amber L. Solberg, *Understanding Large Language Models*, 9 GEO. L. TECH. REV. 256, 262 (2025).

¹⁷⁹ “Max output length” refers to the tokens produced in the model’s response. A longer output is not always more effective output. See Sania Nayab et al., *Concise Thoughts: Impact of Output Length on LLM Reasoning and Cost 1* (Jan. 23, 2025), <https://perma.cc/VJ7X-ZL6E> (“[P]rompt engineering techniques, such as chain-of-thought (CoT), have gained attention for enhancing the explanation and correctness of outputs. However, many models and techniques tend to produce excessively verbose and lengthy answers, leading to issues with both conciseness and generation time.”).

other AI developers, reasoning models are able to transcend pattern recognition and “understand” ambiguity, abstract concepts, and context in ways that foundation models could not.¹⁸⁰ They can adapt to new situations, make inferences that mimic human “common sense,” and generalize knowledge across different domains.¹⁸¹ A key feature of these new models is that they “take time to ‘think’” through complex tasks before responding,¹⁸² producing more accurate outputs and increasing “reasoning” transparency. Paradoxically, as we explain below, this transparency may create additional avenues for defamatory hallucinations.

The leap from “foundation model” to “reasoning model” did not occur overnight. A major step forward in the race to improve AI’s performance occurred in early 2022 when Jason Wei and his co-authors from Google’s Research Team introduced the chain-of-thought prompting technique to elicit step-by-step reasoning in LLMs like GPT-3.¹⁸³ A “chain of thought” is a series of intermediate reasoning steps that roughly mimics the multi-step reasoning that would be performed by a human solving the same problem.¹⁸⁴ Instead of producing an answer in one leap, the model is explicitly guided through a series of intermediate reasoning steps (a “chain of thought”) before reaching its final conclusion. The concept seems deceptively simple: Just as grade-school students are taught how to “break down” a new math concept and then are asked to “show their work,” the LLM is prompted with exemplars of the types of intermediate questions necessary to “reason” through a problem; as such, the LLM is effectively scaffolded by a “few-shot prompt” consisting of an input, chain of thought, and output.¹⁸⁵

Consider the image below: On the left, a standard, “one-shot” prompting method includes a sample question but displays only its input and output. On the

¹⁸⁰ *Reasoning Models*, OPENAI, <https://perma.cc/BY76-DPYG>; *Reasoning Models Don’t Always Say What They Think*, ANTHROPIC (Apr. 3, 2025), <https://perma.cc/M5KR-L9WP>.

¹⁸¹ A team of Apple researchers in June 2025 questioned the thinking capacity of the newest reasoning models. See Shojaei et al., *supra* note 41.

¹⁸² OPENAI, *supra* note 176. For an excellent primer on reasoning models, see Cameron R. Wolfe, *Demystifying Reasoning Models*, DEEP (LEARNING) FOCUS (Feb. 18, 2025), <https://perma.cc/VN27-MBJQ>.

¹⁸³ Jason Wei et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (Jan. 10, 2023), <https://perma.cc/ZZ9P-6DFV>.

¹⁸⁴ *Id.* See also Schwarcz et al., *supra* note 45, at 5.

¹⁸⁵ Wei et al., *supra* note 183, at 2.

right, CoT prompting includes the intermediate steps (i.e., “What questions do we need to ask to solve this problem?”) necessary to reach the final output as a part of the prompt itself. In the image, the blue highlighted section represents the intermediate steps.

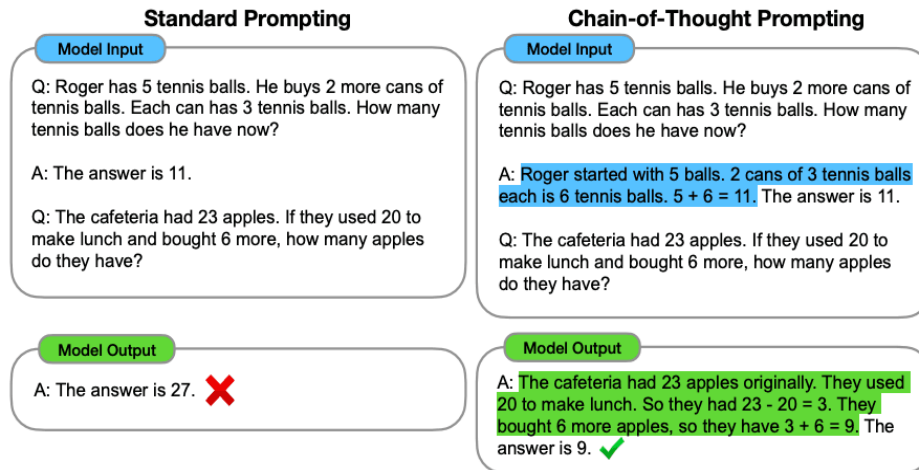


Figure 1: Left, standard prompting. Right, Wei’s CoT prompting.

The advantages of this method were two-fold: First, CoT significantly improved the model’s problem-solving performance on a range of reasoning tasks, making it more adaptable than its predecessors;¹⁸⁶ second, the more detailed model outputs in response to CoT allowed researchers a “window into the behavior of the model.”¹⁸⁷ The transparency of the model’s intermediate steps provides a valuable opportunity for human prompters to understand where the reasoning path has gone wrong when it does, just like a math teacher correcting long-division homework.¹⁸⁸

Four months after Wei and the Google Research team published their findings, another group of researchers led by Takeshi Kojima published their own discovery: Kojima replaced CoT’s intermediate reasoning steps and output with the simple

¹⁸⁶ *Id.* at 1.

¹⁸⁷ *Id.* at 3.

¹⁸⁸ See generally Asifullah Khan et al., *Advances in LLMs with Focus on Reasoning, Adaptability, Efficiency, and Ethics* 5–8 (July 31, 2025), <https://perma.cc/U4KB-YHV4>.

instruction “Let’s think step by step,”¹⁸⁹ and observed increased model accuracy in comparison to standard prompting, albeit less than Wei’s method.¹⁹⁰ Kojima’s method nonetheless had a striking advantage over Wei’s. Wei’s method “hinges on the hand-crafting of task-specific demonstrations one by one,”¹⁹¹ making it too labor-intensive for widespread adoption. Kojima’s reaches similar results by, as he describes it, uncovering the emergent reasoning ability of generative AI.¹⁹² Equally important for CoT’s broad adoption, Kojima’s prompt was universal: It could be used to elicit intermediate reasoning steps across a wide range of tasks using a single template.¹⁹³ And it used a more efficient “zero-shot” prompting method, in which the model “is trained to recognize and categorize objects or concepts without having seen any examples of those categories or concepts beforehand.”¹⁹⁴ A comparison of the two methods can be seen in the figure below, which illustrates the simplicity and universality of the “zero-shot” approach.

¹⁸⁹ Takeshi Kojima et al., *Large Language Models Are Zero-Shot Reasoners* 2 (Jan. 29, 2023), <https://perma.cc/A7MU-PQRK>.

¹⁹⁰ *Id.*

¹⁹¹ Zhuosheng Zhang et al., *Automatic Chain of Thought Prompting in Large Language Models* 1 (Oct. 7, 2022), <https://perma.cc/7UGS-LU39>.

¹⁹² Kojima et al., *supra* note 189, at 2.

¹⁹³ *Id.* at 3.

¹⁹⁴ Compare the few-shot CoT approach, which scaffolds an example of solving a similar problem before providing a new prompt, with a zero-shot approach, which tasks the model with solving a scenario without seeing any examples beforehand. See Dave Bergmann, *What Is Zero-Shot Learning?*, IBM (Jan. 24, 2024), <https://perma.cc/5K4Z-HGXP> (“Zero-shot learning is a machine learning scenario in which an AI model is trained to recognize and categorize objects or concepts without having seen any examples of those categories or concepts beforehand.”).

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Figure 2: On top, Wei's example of Manual-CoT.

On bottom, Kojima's Zero-shot-CoT with universal prompt.¹⁹⁵

Just a few months after the Kojima team made their discovery public, Zhuo-sheng Zhang and a group of Amazon researchers found a “Goldilocks” solution between the universal application of Kojima’s zero-shot CoT and the performance of Wei’s manual CoT.¹⁹⁶ Their approach involved leveraging Kojima’s zero-shot CoT to generate answers that were accompanied by step-by-step explanations and then feeding those explanations into the model as training data (re-creating the function of Wei’s manual prompting).¹⁹⁷

¹⁹⁵ Kojima et al., *supra* note 189, at 2.

¹⁹⁶ Zhang et al., *supra* note 191, at 1.

¹⁹⁷ *Id.* at 6.

Zhang called the process an “automatic chain-of-thought prompting” method, or Auto-CoT.¹⁹⁸ Auto-CoT ensured that high-quality reasoning prompts could be generated at scale without relying on human expertise or labor. The method also incorporated a quality-control step in which multiple reasoning paths were sampled, allowing the model to discard unreliable or inconsistent explanations. The results were striking: “On ten public benchmark reasoning tasks with GPT-3, Auto-CoT consistently match[ed] or exceed[ed] the performance of the CoT paradigm that require[d] manual designs of demonstrations.”¹⁹⁹

But the recent lightspeed pace of model development does not stop there: To date, Chain-of-Verification (CoV) represents the largest step forward for fact-checking in LLM responses.²⁰⁰ The CoV method responds to a flaw in Auto-CoT, namely, that the model might follow a *logical* reasoning process and nonetheless arrive at a *factually incorrect* conclusion. In September 2023, Shehzaad Dhuliawala and colleagues at Meta AI incorporated self-verification mechanisms into the model’s reasoning process²⁰¹ by asking the model to generate and answer fact-checking questions to verify the accuracy of its initial response.²⁰² The answers to these verification questions, which are checked against the model’s baseline response, typically yield higher accuracy than the initial response.²⁰³ Finally, the revised response accounts for the verification questions.²⁰⁴

The figure below illustrates the genius of this method. First, the LLM drafts an initial response to a prompt asking it to name politicians born in New York. Note that the LLM’s first response includes politicians, such as Hillary Clinton, who have a connection to New York but were not born there. Second, the LLM interrogates

¹⁹⁸ *Id.*

¹⁹⁹ *Id.* at 1.

²⁰⁰ Shehzaad Dhuliawala et al., *Chain-of-Verification Reduces Hallucination in Large Language Models* (Sep. 25, 2023), <https://perma.cc/9WUR-5YY4>. We are highly cognizant that by the time this article is published, new developments may have superseded chain-of-verification. Computer scientists continue to seek new methods of improving accuracy and reducing hallucinations. See, e.g., Adarsh Kumar et al., *Improving the Reliability of LLMs: Combining Chain-of-Thought Reasoning and Retrieval-Augmented Generation* (May 13, 2025), <https://perma.cc/JW45-75E6>.

²⁰¹ Dhuliawala et al., *supra* note 200, at 3.

²⁰² *Id.* at 1.

²⁰³ *Id.* at 2.

²⁰⁴ *Id.*

its own answer by generating verification questions, such as “Where was Hillary Clinton born?” Third, the LLM answers the questions independently to ensure the answers are not biased by other responses. Thus, the LLM “learns” that “Hillary Clinton was born in Chicago, Illinois.” Fourth, the LLM generates the verified response and provides it as the final output. Now the LLM correctly excludes Hillary Clinton from the final output because it confirmed, using CoV, that she was *not* born in New York.²⁰⁵

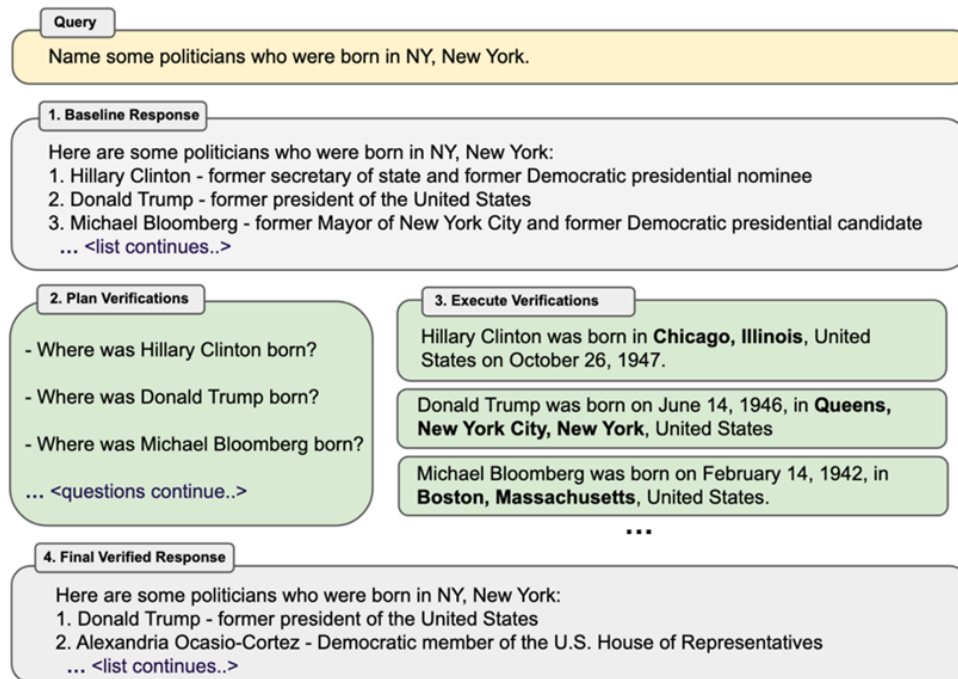


Figure 3: Dhuliawala’s CoV method.

By forcing the model to critically examine its own claims, CoV improves the accuracy and reliability of its final output while also providing transparency and accountability for its decision-making. By doing so, it enacts one of the proposals made by Eugene Volokh’s *Large Libel Models*. Volokh contended that, as a precaution against defamation, “[t]he AI could just automatically recheck its work (at least when some post-processing language recognition suggests that the statement likely contains allegations of misconduct about someone).”²⁰⁶ CoV (as currently implemented in DeepSeek-R1 and OpenAI o1) is exactly what its label implies: a system

²⁰⁵ *Id.*

²⁰⁶ Volokh, *supra* note 26, at 530.

in which an LLM executes piecemeal verifications of its baseline responses before it synthesizes its findings into a final, verified response.²⁰⁷

These scholarly discoveries concerning how to force AI to internalize fact-checking are already incorporated into AI's newest reasoning models. The model used by DeepSeek-R1 looks a bit like AI "brainstorming,"²⁰⁸ as illustrated below. The pictured response resulted from a query made to DeepSeek-R1 by one of the authors of this article. The prompter asked the model to "identify examples of sexual harassment claims against law school professors." DeepSeek first drafted a baseline response. But the baseline response was both wrong and defamatory. The professor named actually worked against sexual harassment, making it easy to see why his name might be proximate to the term in a large data set and why the defamatory hallucination in the initial output might have occurred.²⁰⁹ Unlike a foundation model, a reasoning model does not stop with its initial response, and instead questions itself further. The final output rejects the defamatory chain of thought rather than reproducing it. In essence, this process allows the model to check its own work before "signing off" on the final output.

Another example could be from the University of Chicago Law School. A professor named [REDACTED] was accused of sexual harassment. Wait, was that actually the case? I think there were some allegations, but I'm not entirely sure. Let me verify. Oh, actually, it might be another professor. Alternatively, maybe the case at Vanderbilt Law School with a professor there. Alternatively, I remember a case at the University of Minnesota Law School where a professor was accused. Let me check.

Figure 4: A sample of DeepSeek-R1 using CoV.

Compared to the DeepSeek example, the chain of verification seen in OpenAI's o1-preview model looks much more refined and provides more information to the

²⁰⁷ Dhuliawala et al., *supra* note 200, at 9.

²⁰⁸ See CHATBOT.APP, <https://perma.cc/5R7P-JWEX>.

²⁰⁹ See Bambauer, *supra* note 45, at 345 (discussing how generative AI might misidentify toxic mushrooms as non-toxic based on a similar process).

user about the model's reasoning process.²¹⁰ In designing the o1-preview model, OpenAI "hard-coded" its "double-checking action . . . such that the activity will always run, regardless of whether a user wants it to do so or not."²¹¹ When one of the authors of this article asked OpenAI's o1-preview to give examples of law school professors accused of sexual harassment, it produced the following chain-of-thought before generating its final output:

Examining key law school cases

I'm piecing together sexual harassment cases at notable law schools, like UC Berkeley, Yale, and the University of Illinois. Credible newspaper quotes will be included for accuracy.

Revisiting the selection

Laying out the options for notable law school sexual harassment cases, like John P. Dwyer at UC Berkeley and Alex Kozinski, ensuring credible newspaper coverage is included to support the analysis.

Tackling case selection

I'm focusing on notable sexual harassment cases at Florida International University College of Law and New York Law School, ensuring the examples are credible and relevant for the discussion.

Seeking notable examples

I'm gathering specific sexual harassment cases, like those involving [REDACTED] at UC Hastings and [REDACTED] at the University of Richmond, to provide credible, relevant newspaper quotes.

Highlighting the case

Investigating the 1990s Harvard Law School Title IX lawsuit and Bruce Hay's 2017 allegations to ensure complete understanding.

Figure 5: The OpenAI o1-preview model's response to a query asking it to identify law professors accused of sexual harassment.

²¹⁰ Lance Eliot, *OpenAI's New o1 Model Leverages Chain-of-Thought Double-Checking to Reduce AI Hallucinations and Boost AI Safety*, FORBES (Sep. 15, 2024), <https://perma.cc/2JA4-JPXE>.

²¹¹ *Id.*

Note that two of the three people named in the intermediate chain of verification are completely innocent of the allegations; the authors of this article found no evidence whatsoever that they have been accused of sexual harassment. Indeed, one of them does not even appear to be a real law professor. Thus, there is a defamatory hallucination, but the defamatory hallucination is in the chain of thought and not in the ultimate output. In this instance, it appears that the defamatory hallucination in the chain of thought *may* have helped the model to produce a more accurate final outcome.²¹² This is not clear, however, as some research suggests that the reasoning steps the model displays do not always relate to its ultimate answer.²¹³ Regardless, from a defamation standpoint, reasoning models present new challenges: “As [reasoning models] try to tackle a problem step by step, they run the risk of hallucinating at each step,”²¹⁴ giving rise to more potential avenues for defamation.

Before concluding our survey of the changing landscape of LLMs, it is important to acknowledge the development of another leading approach to enhancing the accuracy and safety of AI outputs: retrieval-augmented generation, or RAG.²¹⁵ To avoid overwhelming our readers with technical details, we have made the main

²¹² Given the accuracy of the final output, this hallucination may be evidence that a chain-of-verification process has worked as intended (i.e., the professors were named in the baseline response, the baseline response was then correctly identified as untrue, the final response excluded the defamatory hallucinations).

²¹³ Yanda Chen et al., *Reasoning Models Don’t Always Say What They Think* 12 (Apr. 3, 2025), <https://perma.cc/Q2GE-UMY2> (“find[ing] that CoT monitoring is a promising approach to noticing unintended behaviors, but that it is not reliable enough to rule out unintended behaviors”). See also *Reasoning Models Don’t Always Say What They Think*, ANTHROPIC (Apr. 3, 2025), <https://perma.cc/7UK9-PQB6> (explaining the Chen et al. findings in language that is easier for those of us who are not computer scientists to understand).

²¹⁴ Cade Metz & Keren Weise, *AI Is Getting More Powerful, but Its Hallucination Are Getting Worse*, N.Y. TIMES (May 5, 2025).

²¹⁵ For a primer on RAG, see Rick Merritt, *What Is Retrieval-Augmented Generation, aka RAG?*, NVIDIA BLOG (Jan. 31, 2025), <https://perma.cc/HK8V-BH4J> (“To understand the latest advancements in generative AI, imagine a courtroom. Judges hear and decide cases based on their general understanding of the law. Sometimes a case—like a malpractice suit or a labor dispute—requires special expertise, so judges send court clerks to a law library, looking for precedents and specific cases they can cite. Like a good judge, large language models (LLMs) can respond to a wide variety of human queries. But to deliver authoritative answers—grounded in specific court proceedings or similar ones—the model needs to be provided that information. The court clerk of AI is a process called retrieval-augmented generation, or RAG for short.”).

focus of this article the application of defamation law to the defamatory outputs of CoT reasoning models. However, much of the analysis is equally applicable to RAG models as well.²¹⁶

When queried, a RAG model not only searches for answers in the database upon which it was trained but also verifies its output by accessing external databases (often provided by the prompter); this “double-checking” allows the model to provide more accurate and up-to-date responses than it otherwise would.²¹⁷ But RAG models also, in some instances, provide the prompting user with links to the sources influencing its conclusions, and the user can employ these links for further fact verification or double-checking.²¹⁸ RAG models are designed to (and do) enhance accuracy relative to prior models.²¹⁹ However, even though RAG models seem to “show their work” in a way that allows for user verification, the unfortunate truth is that RAG models, like CoT models, continue to hallucinate even as they enhance the accuracy of their final outputs.²²⁰ The reasons for this are discussed in the next section.

²¹⁶ See Magesh et al., *supra* note 24, at 1.

²¹⁷ Tianzhe Zhao et al., *RAG Safety: Exploring Knowledge Poisoning Attacks to Retrieval-Augmented Generation* 1 (July 9, 2025), <https://perma.cc/8QPR-EXQL> (“Retrieval-Augmented Generation (RAG) extends Large Language Models (LLMs) with access to external knowledge sources, enabling responses to be grounded in retrieved contents rather than generated solely from the model’s internal parameters. This approach helps mitigate several key limitations of LLMs, such as hallucinations, outdated knowledge, and weak domain adaptation.” (footnotes omitted)).

²¹⁸ Schwarcz et al., *supra* note 45 at 6.

²¹⁹ See Magesh et al., *supra* note 24, at 1 (finding that AI legal research tools hallucinate between 17% and 33% of the time, although they still produce fewer hallucinations than general-purpose chatbots).

²²⁰ *Id.* See also Kyle Wiggers, *Why RAG Won’t Solve Generative AI’s Hallucination Problem*, TECHCRUNCH (May 4, 2024), <https://perma.cc/V7YM-T5GD>. Though beyond the scope of this article, the academy has yet to settle on a canonical definition for precisely which errors are sufficient to constitute a hallucinated response. See Magesh et al., *supra* note 24, at 22 (“A citation, for instance, might point to a case that has been overruled, but that case might still be helpful to an attorney in starting the research process. In our setting, we coded such instances as misgrounded, but whether the model is helpful will still fundamentally have to be determined by use cases and evaluations that involve human interactions with the system.”). Whereas the hallucination in the *Walters* complaint is clearly incompatible with the facts of the world, there exists a “gray area” where outputted responses are not factually *inaccurate* and yet do not fully capture the information contained in the training data.

The rapidity of LLM development creates significant challenges for those tasked with producing effective legal responses.²²¹ First among these is the fact that the pace of change makes it difficult for lawmakers (and scholars!) to keep up with the latest advances. There is a real risk that regulatory responses will be outdated by the time they are implemented, leaving regulators to navigate new terrain with an obsolete map. Second, many of the recent advances in LLM development have arisen as a result of the rapid pace of scholarly discovery: a sign that market incentives to achieve accuracy and reliability are operating at least somewhat effectively.²²² Third, it seems likely that both this rapid pace of change and innovations to improve accuracy will continue, with or without legal intervention.

Fourth, AI is already widespread, and users are becoming increasingly familiar with its capabilities—and its limits.²²³ As users grow more sophisticated in their AI

²²¹ The prompting innovations described above progressing from manual CoT to zero-shot CoT to Auto-CoT each developed inside of one calendar year (2022). The pace of change was so rapid that the Google, Meta, Amazon, and University of Tokyo research teams driving these innovations posted their discoveries online “pre-publication” to a “digital open access,” “curated research-sharing platform,” ArXiv, which is maintained by Cornell University. *About ArXiv*, ARXIV, <https://perma.cc/4UTX-BGUA>. ArXiv “is a free distribution service and an open-access archive for nearly 2.4 million scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics.” *ArXiv.org E-Print Archive*, ARXIV, <https://perma.cc/5AB9-83KD>.

²²² The EU proposes to respond to the rapidity of change in this area by adopting a “regulatory sandbox”—that is, a legal classification that frees AI from some types of regulation for a limited time. The regulatory sandbox approach is purported to “support[] the sharing of best practices through cooperation with the authorities involved in the AI regulatory sandbox.” See Artificial Intelligence Act art. 57, § 6, 2024 O.J. (L 1689) 88 (“Competent authorities shall provide, as appropriate, guidance, supervision and support within the AI regulatory sandbox with a view to identifying risks, in particular to fundamental rights, health and safety, testing, mitigation measures, and their effectiveness in relation to the obligations and requirements of this Regulation and, where relevant, other Union and national law supervised within the sandbox.”). However, query whether this approach can keep pace with the rapid changes described above and whether it is necessary in light of the current system of global, open-source AI scholarship has achieved this sharing of best practices without necessitating the red-tape or delay of government oversight.

²²³ Michelle Faverio & Alec Tyson, *What the Data Says About Americans’ Views of Artificial Intelligence*, PEW RSCH. CTR. (Nov. 21, 2023), <https://perma.cc/WK63-JXS2> (reporting that 90% of Americans know at least “a little” about artificial intelligence, but “[a]dults with a college or post-graduate degree are more likely to be familiar with AI than those with less education,” although “[j]ust 18% of all U.S. adults say they’ve used ChatGPT”).

usage, application of *caveat utilis* becomes a more realistic principle for avoiding some types of harms that stem from incorrect AI outputs. One might counter, though, that, as AI outputs improve in reliability, users are more likely to be lulled into a false sense of security and assume that double-checking AI output is no longer necessary. As AI evolves, the application of liability rules may need to evolve, too.

Fifth, one sometimes must tolerate the bitter with the sweet: The logical structure of CoT reasoning models may simultaneously *amplify* the possibility that defamatory hallucinations will appear in the chain of thought even while it *minimizes* the chance that the final output will be defamatory—a point that is especially crucial for lawmakers or regulators seeking to balance innovation with protection of reputation. That is not to say that real harms to reputation do not deserve compensation (more on that below in Part III.B.1); it is just to say that user awareness about how chatbots operate can help avoid some potential harms, and users themselves may be subject to liability for repeating defamatory outputs without verification. And, sixth and finally, even with the best of intentions and state-of-the-art design, AI models continue to produce defamatory hallucinations. In the next section, we explain why—and why it may not be such a bad thing after all.

B. Hallucinations as Inevitable (and Sometimes Valuable!) Errors

Recent computer science research confirms that LLM hallucinations remain inevitable.²²⁴ Several factors contribute to their ineradicable nature. One is simply

²²⁴ Xu, Jain & Kankanhalli, *supra* note 23, at 2 (showing that “hallucination is inevitable for any computable LLM, regardless of model architecture, learning algorithms, prompting techniques, or training data” (emphasis omitted)). Even setting temperature to “0” will not fully eliminate inconsistent responses. Dahl et al., *supra* note 24, at 67. See also Magesh et al., *supra* note 24, at 1 (finding that “while hallucinations are reduced relative to general-purpose chatbots (GPT-4), . . . the AI research tools made by LexisNexis (Lexis+ AI) and Thomson Reuters (Westlaw AI-Assisted Research and Ask Practical Law AI) each hallucinate between 17% and 33% of the time”); Lauren Leffer, *AI Chatbots Will Never Stop Hallucinating*, SCI. AM. (Apr. 5, 2024), <https://perma.cc/PH6S-BK8E>. One might argue that the term “hallucination” is misleading because it anthropomorphizes a computational operation in a pejorative manner. Likewise, it is something of a misnomer to refer to the hallucination as an “error.” To err is human, and the hallucinatory output is not the product of human thought.

the complexity of language. As in human communication, the receiver of a message (in this case, the AI model) can err in decoding the sender's meaning.²²⁵

Second is the models' underlying structure of algorithmic next-token generation. The design of LLMs relies on probabilistic predictions, and such predictions may easily go astray.²²⁶ A person's name may appear in proximity to a crime because they committed the crime or because they fought the crime;²²⁷ this seems to have been the case in the first-ever defamation action brought against an LLM, which involved an Australian mayor who found himself accused by ChatGPT of bribery, when in fact he had been a whistleblower reporting bribery by others.²²⁸ The important takeaway, though, is that even though AI models are becoming more context-sensitive, they will continue to make errors as long as they rely on probabilistic predictions. Third, hallucinations may arise because the training data fed to a model is inaccurate, biased, conflicting, or contradictory, or because the model's architecture is faulty, causing the AI model to repeat "training patterns rather than generalizing knowledge," or to weigh retrieved information improperly or in a manner that produces incorrect outputs.²²⁹ Finally, AI models are "people

²²⁵ See generally Jiali Cheng & Hadi Amiri, *Linguistic Blind Spots of Large Language Models* (Mar. 25, 2025), <https://perma.cc/4QU5-6GAZ>.

²²⁶ See generally Schwarcz et al., *supra* note 45. But hallucinations occur, see generally Dahl et al., *supra* note 24; Adomas Misiūnas, *Understanding LLM Hallucinations: Causes, Examples, and Strategies for Reliable AI-Generated Content*, CYBERNEWS (Aug. 12, 2025), <https://perma.cc/DK8L-7V63> ("Given that LLMs are stochastic, probability-driven token prediction rather than proper comprehension occurs. Tokenization issues also contribute to hallucinations, as tokenizers break text into numerical tokens that can sometimes be misinterpreted.").

²²⁷ See Bambauer, *supra* note 45, at 345 ("For instance, while a person reading up about mushrooms would be able to easily recognize that their sources make no claims at all about the safety of eating a particular mushroom while noting that others are non-toxic, an AI that uses certain large language model processing might associate the name of the toxic mushroom with the non-toxic notation of other mushrooms that happens to appear nearby.").

²²⁸ David Swan, *Australian Mayor Abandons World-First ChatGPT Lawsuit*, SYDNEY MORNING HERALD (Feb. 12, 2024), <https://perma.cc/826D-MXBR>.

²²⁹ Kromerovas, *supra* note 226. Query whether retrieved information should be weighed descriptively, based on aggregated speech patterns, or normatively, according to a widely accepted set of first principles or values? This topic is one that should be explored in future scholarship.

pleasers”: They often try to fill gaps in their knowledge in order to provide an answer, even when they lack adequate data to generate an accurate one.²³⁰

Paradoxically, however, this novel gap-filling capacity may be essential for some of AI’s more creative uses. Recombining existing ideas or data in novel ways is the very essence of creativity.²³¹ This is what it means to “think outside the box.” The utility of increasingly advanced AI models may thus be inextricably linked with their propensity to hallucinate, and thereby with their associated danger. If so, to eradicate the “inevitable errors” made by LLMs may eradicate a significant part of what makes them valuable, or even functional.²³²

²³⁰ Tor Constantino, *Can You Trust AI Search? New Study Reveals the Shocking Truth*, FORBES (Mar. 28, 2025), <https://perma.cc/79RF-T32L> (reporting results of a new study finding that “Chatbots were generally bad at declining to answer questions they couldn’t answer accurately, offering incorrect or speculative answers instead. Generative search tools often fabricated links and cited syndicated and copied versions of articles.”); Leffer, *supra* note 224 (“LLMs are built to *always* produce an answer, even on topics that don’t appear in their training data.”). A potential release on this “pressure valve” is abstention: A model may choose *not to respond* to certain requests. How and when a model may be trained to abstain from prompt requests is beyond the scope of this paper. For an optimistic outlook on the importance of strategic abstention, see Bingbing Wen et al., *Know Your Limits: A Survey of Abstention in Large Language Models* (Feb. 12, 2025), <https://perma.cc/CS79-JKLZ>.

²³¹ “Creativity is the ability to make or bring into existence something new.” Barbara Kerr, *Creativity*, BRITANNICA (Sep. 1, 2025), <https://perma.cc/V3KE-3G5V>. For an applied example, see Dahl et al., *supra* note 24, at 68 (“As mentioned, the first two modes of hallucination are not always problematic in the legal setting: [T]hese kinds of hallucinations could actually be somewhat desirable to lawyers if they resulted in generated language that, for example, removed unnecessary information from a given argument (at the expense of being faithful to it) or invented a novel analogy never yet proposed (at the expense of being grounded in the lexicon).” (citation omitted)).

²³² See Qiucheng Chen & Bo Wang, *Valuable Hallucinations: Realizable Non-Realistic Propositions 2* (Feb. 18, 2025), <https://perma.cc/C9Q7-2R4B> (“defin[ing] ‘valuable hallucinations’ as realizable but non-realistic propositions. These propositions, if realized, could offer innovative and inspiring ideas, providing new perspectives or solutions to real-world problems.”). Some models allow a user to reduce this type of hallucination by adjusting the “temperature” setting. A reduced temperature setting leads to fewer hallucinations/more fidelity to the training corpus but also to less creativity. Damon Garn, *Understanding the Role of Temperature Settings in AI Output*, TECHTARGET (Mar. 20, 2025), <https://perma.cc/8DNJ-J67P>.

LLM hallucinations come in various forms.²³³ In fact, a team of researchers at Yale recently created a taxonomy of hallucinations, dividing them into (1) responses inconsistent with or unfaithful to the user's prompt; (2) responses inconsistent with the model's training corpus; and (3) responses inconsistent with the facts of the world.²³⁴ Each raises different concerns for defamation.

The first type occurs when an AI model has access to all the data needed to answer a question but mischaracterizes the data.²³⁵ For example, if a user asks o1 to summarize a legal document, provides the entire legal document in the prompt, and o1 then outputs a response that mischaracterizes one of the witnesses to a crime as the alleged perpetrator, that output is not consistent with the prompt and may be defamatory. However, the user who submitted the prompt presumably has the document at hand to double-check the work of the AI, and the user's unthinking reliance on the output is unjustified.²³⁶ Or, at least, it is unjustified when anything significant hinges on the model's output being correct!

The second type of hallucination occurs when an AI model produces an output that is not "logically derivable from the content of its training corpus," whether or not the information in that training corpus is itself factually true.²³⁷ Dahl and his co-authors provide a clever example:

[I]f a training corpus consisted of J. K. Rowling's *Harry Potter* series, we would expect an LLM to produce the sentence "Tom Marvolo Riddle" in response to a query about

²³³ Samuel Greengard, *Shining a Light on AI Hallucinations*, COMM'NS ACM (Apr. 3, 2025), <https://perma.cc/T2LU-SQ7A> (discussing the value of hallucinations).

²³⁴ See Dahl et al., *supra* note 24, at 67.

²³⁵ This is known as a closed-domain or intrinsic hallucination. See *id.* It is called "closed domain" because the LLM is drawing from a closed universe of information to provide the output to the prompt.

²³⁶ For this reason, legal AI professionals have advocated for a "trust but verify" approach to even those Legal AI tools which promise virtually hallucination-free outputs. Ideally, a user would "fine-tune" the model, via temperature settings or custom system instructions, to reflect the level of discretion a model uses in "paring down" a legal document in a closed-domain setting. Geoffrey D. Ivnik, *Trust Me I'm a Legal AI: Can the Legal Profession Close the 'Trust Gap' with Gen AI?*, LEXISNEXIS (July 10, 2024), <https://perma.cc/BP4S-XKMY>.

²³⁷ Dahl et al., *supra* note 24, at 67. See generally Ayush Agrawal et al., *Do Language Models Know When They're Hallucinating References?*, 2024 FINDINGS ASS'N FOR COMPUTATIONAL LINGUISTICS: EACL 912.

Voldemort’s real name. However, if the training corpus consisted solely of Jane Austen’s *Pride and Prejudice* (for instance), we would consider this LLM output to be a hallucination—because there would be no basis in the training data for making such a claim about Voldemort.²³⁸

Although this example might lead a user to believe there was a defect in the model’s training data, similar “outside the training-data box” examples might prove quite valuable to users and to society. LLMs are sometimes able to provide innovative and original solutions to complex problems precisely because their capacity to synthesize and recombine information exceeds that of humans, and they simply “think differently” than humans.²³⁹ The reasons a fine-tuned LLM may sometimes “invent[] a novel analogy never yet proposed”²⁴⁰ has to do with the way they draw from massive amounts of data²⁴¹ and represent words through a list of “word vectors” rather than a sequence of letters (like humans do).²⁴² Some models even allow users to adjust the probability distribution of potential next words or tokens in order to enhance the potential for creative outputs; adjusting a model to a higher “temperature” setting means that the model is more likely to combine less probable words or tokens in its output. This setting may be desirable when one wishes to encourage creative outputs, but it also increases the risk of incoherent or factually incorrect outputs.²⁴³ Even at low temperature settings, however, hallucinations remain, and the results produced sometimes defy human explanation because of the “black box

²³⁸ Dahl et al., *supra* note 24, at 67 n.3.

²³⁹ See *id.* at 68 (“At the same time, however, insofar as creativity is valued, certain legal tasks—such as persuasive argumentation—might actually benefit from some lack of strict fidelity to the training corpus; after all, a model that simply parrots exactly the text that it has been trained on could itself be undesirable.” (emphasis omitted)).

²⁴⁰ *Id.*

²⁴¹ STEFAN BAACK, MOZILLA, TRAINING DATA FOR THE PRICE OF A SANDWICH: COMMON CRAWL’S IMPACT ON GENERATIVE AI (2024), <https://perma.cc/8YJG-FXN2>.

²⁴² For a primer on next-token prediction in comparison to human language, see Lee & Trott, *supra* note 9.

²⁴³ Dahl’s team performed their experiments with the model temperature set to “0” and still observed hallucinations. See Dahl et al., *supra* note 24, at 67. For an example of using generative AI to “brainstorm” original metaphors or kennings, try asking it to provide kennings similar to “whale-road” (describing the sea) and see how many similar kennings it can generate. Examples may include “star-blanket” and “cloud-meadow” for the sky, “night-lantern” or “tide-puller” for the moon, and “wood-eater” or “hearth-blossom” for fire.

problem” described above. Our current inability to explain how deep learning systems reach some of their conclusions²⁴⁴ is one of the reasons that hallucinations of all three types are here to stay.

The third and final type of hallucination is the one that most concerns defamation law, because it is the one most likely to cause reputational harm. This type occurs when the model’s response is “inconsisten[t] with the facts of the world.”²⁴⁵ This type is the classic “hallucination” discussed in media accounts, such as those involving lawyers submitting non-existent legal cases to bemused and angry judges.²⁴⁶ It is also the type of hallucination produced in the *Walters* case. Although GPT-3.5 initially signaled that it lacked the capacity to answer the prompter’s query,²⁴⁷ it produced, when prodded, an entirely fabricated document by apparently extrapolating from the facts given in the prompt and the information it had in its training data about related matters.

The *Walters* hallucination, and the hallucinations produced by our prompts about law professors accused of sexual harassment, reveal relevant but perhaps uncomfortable truths about potentially defamatory hallucinations. First, LLMs are more likely to produce hallucinations about the most vulnerable category of potential defamation plaintiffs—private figures. Because LLM data sources are less likely to contain information about private figures, LLMs have less ability to cross-check potential outputs, and are thus more likely to simply hallucinate about private figures than they are about public ones.

Second, and relatedly, user behavior affects the likelihood of hallucinations. If the *Walters* facts are any guide, it appears that LLMs may be more prone to hallucinate when prodded repeatedly or asked for evidence. And certainly user prompts are also more likely to generate hallucinations if they are not written clearly, if they provide little context, or if they take the form of leading questions, such as “What

²⁴⁴ Lou Blouin, *AI’s Mysterious ‘Black Box’ Problem, Explained*, UNIV. MICH.-DEARBORN NEWS (Mar. 6, 2023), <https://perma.cc/XG43-2ZFL>.

²⁴⁵ Dahl et al., *supra* note 24, at 67.

²⁴⁶ See *Mata v. Avianca*, 678 F. Supp. 3d 443 (S.D.N.Y. 2023) (involving lawyer who submitted non-existent cases in court); Complaint, *supra* note 27 (defamed party accused of crimes); Byron Kaye, *Australian Mayor Readies World’s First Defamation Lawsuit Over ChatGPT Content*, REUTERS (Apr. 5, 2023), <https://perma.cc/A479-P8LS>.

²⁴⁷ Complaint, *supra* note 27, at 3, ¶ 25.

crime did Lyrissa Lidsky commit?”²⁴⁸ Similarly, a user’s failure to provide feedback when the user knows a model has provided them with erroneous information may make future hallucinations more likely, given the iterative nature of model design.²⁴⁹ A user can even influence the likelihood of hallucinations by using a model unsuited for the user’s research task.²⁵⁰ If AI is truly to become a tool for everyone, AI producers will have to improve how models respond to flawed prompts as well as educate users about their roles in improving model outputs and reducing undesirable hallucinations.

A third, possibly inconvenient truth is that human users who do not fully understand how LLMs operate (and even some who do) may be prone to granting LLMs Delphic-like authority. “[G]enerative AI outputs . . . create an illusion of credibility—especially when the information is completely made up.”²⁵¹ Because LLMs respond like humans in some respects, humans may have a tendency to trust model outputs more than they should, and this is particularly true for “vulnerable individuals,” for whom LLMs may “reinforce delusional thinking” because of their sycophantic design.²⁵² It is an open question whether the dangers of anthropomorphizing AI will diminish as users and LLMs become more sophisticated. However, new research suggests that certain design choices may reduce the risk of users anthropomorphizing chatbots.²⁵³ If so, such design choices should be encouraged and perhaps, at some future point, mandated.²⁵⁴

²⁴⁸ Volokh, *supra* note 26.

²⁴⁹ See generally Nathan Lambert, *Reinforcement Learning from Human Feedback*, <https://perma.cc/6NYU-Z8MK>.

²⁵⁰ For example, a domain-specific LLM, such as one trained on medical data, might be better suited to predict the correct medication dosage for a particular patient than a general purpose LLM.

²⁵¹ Klaudia Jąźwińska & Aisvarya Chandrasekar, *AI Search Has a Citation Problem*, COLUM. JOURNALISM REV. (Mar. 6, 2025), <https://perma.cc/V9DZ-PAHX>; Constantino, *supra* note 230.

²⁵² Kashmir Hill, *They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling.*, N.Y. TIMES (June 13, 2025) (documenting cases of vulnerable individuals being drawn into false “conspiracy theories” generated by AI); Ali Hasan, *Are You Anthropomorphizing AI?*, BLOG APA (Aug. 20, 2024), <https://perma.cc/5SKG-NWS4>.

²⁵³ Myra Cheng et al., *Dehumanizing Machines: Mitigating Anthropomorphic Behaviors in Text Generation Systems* (June 3, 2025), <https://perma.cc/RA25-WBGN>. See also *Early Methods for Studying Affective Use and Emotional Well-Being on ChatGPT*, OPENAI (Mar. 21, 2025), <https://perma.cc/S66X-QKV6>.

²⁵⁴ This topic, while fascinating, lies outside the scope of this paper.

III. APPLYING DEFAMATION LAW TO LLM REASONING MODELS: EVALUATING POLICY TRADEOFFS AND EXISTING PROPOSALS

Depending on the public policy choices made by courts and legislators at this stage of LLM development, developers such as OpenAI may be incentivized to make unfortunate tradeoffs that affect future transparency and innovation in the operation and design of their models.²⁵⁵ If, for example, defamation law incentivizes model developers to hide chains of thought, innovation to produce better outputs may suffer in both the short and long terms. Conversely, law might be used to incentivize the retention of records allowing those defamed to document the scope of reputational harm they suffer. In this Part, we underscore six insights gleaned from our discussion of the changing LLM landscape and we identify the policy goals that should underpin the adaptation of common-law tort doctrines to defamation-by-hallucination cases arising from the new reasoning models.

A. *Six Insights from the Changing LLM Landscape*

Our survey of LLM landscape developments over the last 24 months reveals six interconnected insights that should inform legal policymakers tasked with developing intelligent communications policy around AI, whether through application of common-law liability rules or development of statutory or regulatory frameworks.

First, AI has developed at an extraordinary pace since the *Walters* case was filed, and this pace seems likely to continue for the foreseeable future. This fact should encourage regulatory caution, lest intervention chill innovation.²⁵⁶

²⁵⁵ Censoring CoTs to eliminate defamatory hallucinations may result in less effective CoT monitoring, undermine the function of CoTs, and hinder the development of safer and more efficient models. See ANTHROPIC, CLAUDE 3.7 SONNET SYSTEM CARD 15–18 (2025), <https://perma.cc/4DAA-S9FC> (“For CoT monitoring to be most effective, the CoT must be a faithful and complete reflection of the way the model reached its conclusion and generated a user-facing response. This means that the model’s CoT must highlight the key factors and steps behind its reasoning. If CoT is not fully faithful, then we cannot depend on our ability to monitor CoT in order to detect misaligned behaviors, because there may be important factors affecting model behavior that have not been explicitly verbalized. . . . The model’s thought process can reveal alignment issues that may not always be evident in the response alone.”).

²⁵⁶ See *Cubby, Inc. v. CompuServe Inc.*, 776 F. Supp. 135, 139 (S.D.N.Y. 1991); *Stratton Oakmont, Inc. v. Prodigy Servs. Co.*, 1995 WL 323710, at *4 (N.Y. Sup. Ct. May 24, 1995) (exemplifying

Second, many of the changes in the AI industry over the last two years have occurred in response to market pressures for producers to increase model accuracy and safety—though the path to improvement has not been linear.²⁵⁷ Both retrieval-augmented generation and chain-of-thought reasoning were developed to satisfy user demand for content verification. For example, retrieval-augmented generation “documents” the accuracy of its output with citations or links to sources, which allows users to trace information to its source within AI training data. Meanwhile, as noted above, chain-of-thought models have a built-in and often transparent process to “double-check” their own work. And these innovations have led to improvement in some cases: Some reasoning models produce more factual outputs than their foundation model predecessors,²⁵⁸ and new discoveries designed to increase model accuracy and safety are rolling out almost weekly,²⁵⁹ though the current state of progress is not necessarily reassuring.²⁶⁰

Third, design transparency supports further innovations.²⁶¹ Some accuracy and safety gains have resulted from open-source AI research. Researchers are racing to share their findings—even before official publication—and this research competition and sharing should, ultimately, propel safety advances in LLM design.²⁶²

courts’ early adaptation of defamation law to online intermediaries, which spurred Congress to enact section 230 of the Communications Decency Act, 47 U.S.C. § 230, though that proved to be an imperfect legislative solution).

²⁵⁷ Magesh et al., *supra* note 24, at 1; *but see* Metz & Weise, *supra* note 214 (reporting that reasoning models are “generating more errors, not fewer” and stating that OpenAI “found that o3—its most powerful system—hallucinated 33 percent of the time when running its PersonQA benchmark test, which involves answering questions about public figures. That is more than twice the hallucination rate of OpenAI’s previous reasoning system, called o1.”).

²⁵⁸ Adam Zewe, *Researchers Reduce Bias in AI Models While Preserving or Improving Accuracy*, IDSS MIT (Dec. 13, 2024), <https://perma.cc/9JWT-L5HT>. *But see* Richard Ren et al., *Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?* (Dec. 27, 2024), <https://perma.cc/6NDL-U5Q3>.

²⁵⁹ Xuli Tang et al., *The Pace of Artificial Intelligence Innovations: Speed, Talent, and Trial-and-Error*, 14 J. INFORMETRICS 101094, at 1 (2020).

²⁶⁰ Metz & Weise, *supra* note 214 (reporting that some new reasoning models had “hallucination rates” “as high as 79 percent” on one test).

²⁶¹ Nagadivya Balasubramaniam et al., *Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements*, 159 INFO. & SOFTWARE TECH. 107197, at 1 (2023).

²⁶² Magesh et al., *supra* note 24.

Fourth, for the reasons explained in the previous section, hallucinations are a feature rather than a bug: they are ineradicable, sometimes inexplicable, and sometimes valuable. Further, the number of hallucinations may continue to diminish but is unlikely to ever reach zero.²⁶³

Fifth, LLMs are capable of producing records of their own malfeasance. Such records can allow decision-makers to more accurately determine how much, if at all, a hallucination harmed its subject.

Finally, as generative AI users become more familiar with AI, they will be unable to reasonably claim to believe that LLM outputs are invariably accurate. Although the various lawyers who have submitted LLM-fabricated case law to courts learned the hard way that AI hallucinates,²⁶⁴ most users of LLMs should soon be aware, if they are not already, that it is unreasonable to rely on a model's output without verification of the information contained therein. AI producers, of course, play a role in shaping user expectations and behaviors: Indeed, they owe it to users to educate them about their roles in reducing AI risks. Regardless, as users become more familiar with AI's inevitable errors, it may eventually be possible for them to reduce the likelihood that those errors occur or that they cause harm when they do.

Generative AI may be transformative, but LLMs are not infallible oracles. Taking these considerations into account helps inform the policies we should strive to achieve in applying defamation law to the newest reasoning models.

B. Communications Policy and the Defamation-by-Hallucination Problem

Defamation law reflects society's "basic concept of the essential dignity and worth of every human being."²⁶⁵ Ideally, defamation law protects reputation by vindicating reputational injury and compensating the dignitary, relational, and economic harms that defamed individuals suffer. Defamation law enforces social norms about what types of speech exceed the limits of civilized discourse,²⁶⁶ but it also creates carve-outs (whether through common-law development, constitu-

²⁶³ Xu, Jain & Kankanhalli, *supra* note 23.

²⁶⁴ See, e.g., Olivia Prentzel, *MyPillow CEO's Lawyers Fined for AI-Generated Court Filing in Denver Defamation Case*, COLO. SUN (July 7, 2025), <https://perma.cc/GK5X-UMAX>.

²⁶⁵ *Rosenblatt v. Baer*, 375 U.S. 75, 92 (1966) (Stewart, J., concurring).

²⁶⁶ For further explanation, see Lyrissa Lidsky, *Untangling Defamation Law: Guideposts for Reform*, 88 MO. L. REV. 663, 677 (2023).

tional interpretation, or statutory enactment) to safeguard the free flow of information, particularly information that enables informed decision-making, participation in democratic processes, or participation in the formation of public opinion.²⁶⁷

In addressing defamation by hallucination, at least at this stage in the development history of LLMs, defamation doctrines should balance the need to compensate individuals who suffer real reputational harm with the need to accommodate those inevitable errors in AI models that are inextricably linked with their creativity and innovation. In striking this balance, the policy goals below should be at the forefront. In assessing these goals, three caveats are in order.

First, this discussion of policy goals takes place at a relatively high level of generality, and defamation law will not be the optimal tool for advancing all of them. For example, some goals may be best advanced by statutes or other bodies of law or regulation. Second, the analysis here is not intended to be a guide for the development or application of legal rules or doctrines to high-risk AI systems, such as weapons or health devices. Instead, this discussion about policy objectives is centered around the problems of preventing, mitigating, and compensating reputational harms resulting from defamatory LLM outputs. Third, these recommended policy goals exist in some tension with one another, which means that the ultimate balance struck will inevitably reflect normative judgments about the goals' relative importance.

1. Compensate defamation victims for provable harms

In many jurisdictions, defamation law allows plaintiffs to recover presumed damages—damages awarded without evidence of actual loss—in some types of cases. The reason? Real reputational harms, such as damages to human relationships, can occur through subtle means not susceptible of easy proof, as discussed above.²⁶⁸ But presumed damages are a second-best substitute designed to achieve the law's goals of compensation and vindication for defamation victims.

²⁶⁷ See generally *id.*

²⁶⁸ As Yonathan Arbel points out, one role of presumed damages may be to compensate harms related to the lowering of one's social status. Yonathan A. Arbel, *A Social Status Theory of Defamation Law*, 14 U.C. IRVINE L. REV. 768, 789 n.145 (2024). The Supreme Court does not allow the award of presumed damages for the violation of abstract constitutional rights, stating, "were such damages

In the context of generative AI, the law may not have to settle for second-best. If, as the *Walters* case and other evidence²⁶⁹ suggests, LLMs can keep a record of how many times a defamatory hallucination about an individual is produced and to whom, such a record may serve as evidence of the likely extent of that hallucination's real harms or a rebuttal of a presumption of harm.²⁷⁰ In addition, such a record might highlight the effect the user's prompts had in leading the model to generate the allegedly harmful hallucination. Finally, as the complaint in the *LTL* case suggests, such a record might help a defamed company prove definitively why a customer has cancelled a contract or taken their business to a competitor. Going forward, the degree to which "chatlogs" assist plaintiffs in proving concrete harms in tort cases should arguably prompt policymakers to require LLMs to keep such records, at least for a reasonable length of time.²⁷¹ Such a record-keeping requirement would help advance the compensatory goal of tort law by allowing the plaintiff to pinpoint harm instead of relying on a presumption.

Although the prospect of having to pay presumed damages might deter human speakers from negligently inflicting reputational harm, this prospect will not deter an LLM from hallucinating. And although the imposition of liability may affect subsequent design choices, these choices will come with other tradeoffs. We explore these tradeoffs in the next section.

available, juries would be free to award arbitrary amounts without any evidentiary basis, or to use their unbounded discretion to punish unpopular defendants." *Memphis Cmty. Sch. Dist. v. Stachura*, 477 U.S. 299, 310 (1986). The Court has, however, allowed presumed damages as a remedy for depriving a voter of the right to vote because "each member of the jury has personal knowledge of the value of the right." *Id.* at 311 n.14.

²⁶⁹ Constantino, *supra* note 230.

²⁷⁰ To fully flesh out our uncertainty, Riehl, the reporter in the *Walters* case, received copies of his ChatGPT logs when he requested them via email from OpenAI.

²⁷¹ See *How to Delete and Archive Chats in ChatGPT*, OPENAI, <https://perma.cc/K4FA-VP8F> (specifying OpenAI's current policy on chat deletion and retention for first-party requests). OpenAI's policy regarding third-party requests for chats is less clear. For example, could OpenAI conduct a search for any time a private individual's name has been generated in response to a user prompt? If so, may OpenAI access chats from even those accounts who have "opted out" of providing their chats for training data?

2. Correct the record and prevent future errors

Legal rules should incentivize the retraction and correction of defamatory falsehoods where feasible. In traditional defamation law, public retraction of a defamatory falsehood helps vindicate a plaintiff's reputation and mitigate damages. Retraction is a "restorative remedy" designed to prevent future harm to the plaintiff,²⁷² but it is also a form of recantation that aligns with slander's origins as a sin punishable in ecclesiastical courts.²⁷³

However, if a defamatory hallucination only reached a tiny audience in the first instance, public retraction may do more harm than good by bringing increased attention.²⁷⁴ If a reasoning model produces a defamatory hallucination, legal rules should encourage retraction, but perhaps not public retraction, since in many cases the defamatory hallucination's dissemination may not have been widespread. A more tailored type of retraction may be possible: If records are available of which users received the defamatory hallucination, rules could encourage receiver-specific retraction. Moreover, to the extent feasible, legal rules could make it easier for users to report defamatory or false outputs²⁷⁵ and require model developers to correct those errors—though, as discussed in the next section, it is less than clear that "takedown" specifically tailored to an individual defamatory hallucination is possible at this point in AI's development.²⁷⁶

²⁷² Alfred C. Yen, *It's Not That Simple: An Unnecessary Elimination of Strict Liability and Presumed Damages*, 23 HARV. C.R.-C.L. L. REV. 593, 608 (1988).

²⁷³ See Colin Rhys Lovell, *The "Reception" of Defamation by the Common Law*, 15 VAND. L. REV. 1051, 1054–55 (1962) ("[C]anon law considered [slander] to be a sin, demanding penance before there could be absolution of the sinner.").

²⁷⁴ See Sue Curry Jansen & Brian Martin, *The Streisand Effect and Censorship Backfire*, 9 INT'L J. COMMUN. 656, 656 (2015) ("Barbra Streisand's attempt to restrict online views of her residence on a public website had the paradoxical effect of leading to many more views than if she had done nothing. Subsequently, attempts at censorship that end up being counterproductive have been dubbed the 'Streisand effect.'").

²⁷⁵ The law, for example, might require AI model producers to provide a readily accessible "complaint system" for their users. The law might also impose reporting requirements on AI producers about the number and nature of complaints received.

²⁷⁶ It goes without saying, perhaps, that users of LLMs should be incentivized to report defamatory outputs to those who can correct them.

3. Incentivize model transparency to reduce future harms

Legal rules should incentivize model developers to provide transparency to users²⁷⁷ regarding, at a minimum, (1) the sources from which the models derive their outputs; (2) the methods by which they produce such outputs, including moderation; and (3) known limitations in their ability to provide correct answers. Retrieval-augmented generation is directly responsive to the first dimension of transparency via its provision of links to the model's sources, though other AI models have not yet been as forthright in enabling user verification of output veracity.²⁷⁸

Regardless, legal rules should support further innovations in source transparency.²⁷⁹ They should also require transparency regarding AI decision-making and the tradeoffs made by AI producers²⁸⁰ in designing new models. This type of transparency often can improve accuracy, mitigate bias, and build trust in AI outcomes. Transparency achieves this by making the internal workings of AI models more visible and understandable to humans. Transparency is essential for correcting AI

²⁷⁷ The text here discusses user transparency, but other kinds of transparency are also beneficial. Open-source development should also be encouraged, because it helps guarantee that AI systems as a whole are transparent, auditable, and modifiable. See generally Yash Raj Shrestha, Georg von Krogh & Stefan Feurriegel, *Building Open-Source AI*, 3 NATURE COMPUTATIONAL SCI. 908 (2023) (discussing the benefits of open-source AI). Further, although it is partially outside the scope of this article, legal rules should also encourage transparency regarding the training data used by the model. See Beth Stackpole, *Bringing Transparency to the Data Used to Train Artificial Intelligence*, MIT SLOAN (Mar. 3, 2025), <https://perma.cc/QHN7-D9PH>.

²⁷⁸ A March 2025 study by Klaudia Jaźwińska and Aisvarya Chandrasekar studied eight AI search engines and found a lack of transparency regarding which content they use to provide outputs in response to user prompts. As they describe the results of their study, “chatbots’ conversational outputs often obfuscate serious underlying issues with information quality.” Jaźwińska & Chandrasekar, *supra* note 251. Moreover, “premium models” of AI search tools “provided more confidently incorrect answers than their free counterparts.” *Id.* However, the study methodology limits the conclusions that can be drawn from it. Because the authors were concerned about AI attribution to original news sources, their chosen methodology was to copy a block quote into an LLM and ask it to retrieve the original article. Responses did not always link to the original source.

²⁷⁹ Rebecca Aviel et al., *From Gods to Google*, 134 YALE L.J. 1269 (2025) (discussing constitutional limitations that may hamper legislative attempts to impose transparency requirements).

²⁸⁰ We have chosen to use the term “producers” as a shorthand covering the various actors in the chain of “production” of an LLM such as ChatGPT, including, but not limited to, creators, data scientists who develop models, dataset generators, data preprocessors, trainers, evaluators, engineers, deployers, monitors, and modifiers.

errors because it can reveal flawed training data, overreliance on particular types of data, or other shortcomings in the algorithm's reasoning. Transparency is also critical for AI's legitimacy; people need to know that high-stakes decisions affecting their lives, such as loan approvals and medical diagnoses, are made fairly and based on relevant factors.

LLM producers have already been transparent regarding the tradeoffs they have made in deploying chain-of-thought reasoning models. OpenAI, for example, explained the tradeoffs it made in deciding which portions of CoT to make available to users in the "system card"²⁸¹ accompanying o1's release. OpenAI acknowledged that the model's choices might sometimes produce the possibility of a discrepancy between the published chain of thought and the model's actual reasoning,²⁸² but said it was choosing to hide "raw" CoT from users in favor of "teaching the model to reproduce any useful ideas from the chain of thought in the answer."²⁸³ Other AI producers have been similarly transparent but have made different design choices and different tradeoffs—choosing to provide users with "warts and all" CoT outputs that show the errors their models are making.²⁸⁴ Such transparency allows for

²⁸¹ A system card lays out the safety measures undertaken by an AI producer prior to releasing its model to the public. See *OpenAI o1 System Card*, OPENAI (Dec. 5, 2024), <https://perma.cc/A3PE-9AHM> ("This report outlines the safety work carried out prior to releasing OpenAI o1, including external red teaming and frontier risk evaluations according to our Preparedness Framework.").

²⁸² It did so partly because "raw" CoT data could not be trained to comply with policy or reflect user preference. In other words, OpenAI transparently chose *not* to provide users with all components of the actual chain of thought "after weighing multiple factors including user experience, competitive advantage, and the option to pursue the chain of thought monitoring." *Id.* ("Until now, that latent thinking has only been available in the form of activations—large blocks of illegible numbers from which we have only been able to extract simple concepts. Chains-of-thought are far more legible by default and could allow us to monitor our models for far more complex behavior (if they accurately reflect the model's thinking, an open research question)." (footnotes omitted)).

²⁸³ OpenAI has also provided its "Model Spec" to the public, which specifies OpenAI's approach to "shaping desired model behavior and how [it] evaluate[s] tradeoffs when conflicts arise." *Introducing the Model Spec*, OPENAI (May 8, 2024), <https://perma.cc/SR46-SXLA>. For example, one of OpenAI's model specs provides: "Don't try to change anyone's mind," a spec that obviously represents a tradeoff between factual accuracy and the psychological comfort of the user. *Id.*

²⁸⁴ In a study of early machine-learning models used by doctors in the diagnosis and treatment of diseases, Agarwal, Tanneru, and Lakkaraju observed that "unfaithful explanations have huge negative consequences leading to incorrect treatment plans and patient harm when the doctor accepts

the possibility of public participation in influencing future tradeoffs and helps educate the public about the “unavoidably unsafe” aspects of LLM usage. It also helps foster iterative improvements in AI development that may produce more innovations for safety in the long term.²⁸⁵

Finally, AI producers need to condition users to appreciate that false-yet-plausible outputs are likely to occur and can be harmful. Thus, as other scholars have argued, legal rules should create a duty on the part of AI producers to warn users not to rely on AI outputs without verification, and the more specific the warning is about the model’s limitations in responding to a user’s actual query, the better.²⁸⁶

4. Incentivize innovations for safety and accuracy

Legal rules should encourage further AI innovation, lest we strangle technological development in its infancy. AI developers or providers should not be held liable solely for failure to eliminate defamatory hallucinations, especially where hallucinations remain inevitable.²⁸⁷

More importantly, legal rules should not chill experimentation that may foster accuracy and safety in AI models or otherwise benefit users and society in the long run. While powerful market incentives²⁸⁸ currently are exerting a powerful influence on AI producers to increase model accuracy, not all market incentives align

explanations without any sanity checks.” Chirag Agarwal, Sree Harsha Tanneru & Himabindu Lakkaraju, *Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models* 5 (Mar. 14, 2024), <https://perma.cc/2XKJ-L9LP>.

²⁸⁵ Caution is necessary in encouraging pseudo-transparency, and policymakers must be attentive to the possibility that the AI model’s stated chain of thought is not “faithful” to its actual decision-making process. See generally Iván Arcuschin et al., *Chain-of-Thought Reasoning in the Wild Is Not Always Faithful* (June 17, 2025), <https://perma.cc/YUN9-6TWR>.

²⁸⁶ Ayres and Balkin argue for such a duty, but they also contend that merely warning users of the possibility of defamatory hallucinations is insufficient to negate the liability of the model provider. See Ayres & Balkin, *supra* note 34.

²⁸⁷ In the early days of airplanes, a plane crash could not, standing alone, provide evidence of negligence via *res ipsa loquitur*. Today, it can. Perhaps one day we will reach a similar stage in AI development. But we’re not there yet.

²⁸⁸ These include capturing market share, increasing government and venture capital funding, increasing brand recognition, creating user loyalty, protecting intellectual property, and reducing development costs. Many of these interests are furthered by enhancing the accuracy and reliability of model outputs. See MIT TECH. REV. INSIGHTS & DATABRICKS, *THE GREAT ACCELERATION: CIO PERSPECTIVES ON GENERATIVE AI* (2023), <https://perma.cc/6SHH-YTVQ>.

with the interests of users and society at large.²⁸⁹ If legal rules are to foster innovation going forward, they must be consistent, predictable, and stable while remaining flexible enough²⁹⁰ (or amendable enough) to adapt to the AI industry's rapid pace of change.²⁹¹ Given the nature of the risks posed by AI, however, this rapid pace of change also warrants mandatory oversight and auditing on an ongoing basis.

5. Incentivize the exercise of reasonable care by AI producers

Humans create, train, and use AI models²⁹² and, as a general proposition, should exercise reasonable care for the protection of others when doing so.²⁹³ Although tort law treats negligently designed information products²⁹⁴ differently than negligently designed products that pose a risk of physical harm, the principle holds, nonetheless, that the law should incentivize LLM developers to reasonably balance risks and utility in creating, designing, and training their AI products in order to avoid causing reputational harm.

²⁸⁹ See generally KATE CRAWFORD, *ATLAS OF AI: POWER, POLITICS, AND THE PLANETARY COSTS OF ARTIFICIAL INTELLIGENCE* 15 (2021) (discussing AI as an "extractive industry" that exploits individuals' data, environmental resources, labor, and energy).

²⁹⁰ The EU has embraced the idea of a "regulatory sandbox at [a] national level." See Artificial Intelligence Act art. 57, § 1, 2024 O.J. (L 1689) 88. A regulatory sandbox is a closed, controlled environment in which AI companies can test their new models under the supervision of regulators but with some regulations relaxed during the period of experimentation. "Member States shall ensure that their competent authorities establish at least one AI regulatory sandbox at national level, which shall be operational by 2 August 2026. That sandbox may also be established jointly with the competent authorities of other Member States. The Commission may provide technical support, advice and tools for the establishment and operation of AI regulatory sandboxes." *Id.* art. 57, § 1.

²⁹¹ S. Shyam Sundar, Cason Schmit & John Villasenor, *Regulating AI: 3 Experts Explain Why It's Difficult to Do and Important to Get Right*, CONVERSATION (Apr. 3, 2023), <https://perma.cc/RU7W-494Z>.

²⁹² See Emily Rumick, *What Happens When Robots Lie? Combatting the Harmful Threats of AI-Generated Disinformation While Harnessing Its Potential*, 25 J.L. SOC'Y 146, 174 (2025).

²⁹³ Ayres & Balkin, *supra* note 34 ("[D]esigners of generative AI systems should have a duty to implement safeguards that reasonably reduce the risk of producing defamatory content. This duty includes a duty of reasonable care in choosing materials for pre-training and fine tuning. It also includes a duty to design and incorporate algorithms that can detect and filter out potentially harmful material, a duty to conduct thorough testing to identify and mitigate risks, and a duty to continually update systems in response to new problems and threats.").

²⁹⁴ Deana Pollard Sacks, *Constitutionalized Negligence*, 89 WASH. U. L. REV. 1065 (2012).

6. Center liability on the cheapest cost avoider

In tort law, the “cheapest cost avoider” is the party who can most easily and cost-effectively avoid wrongfully inflicting the harm.²⁹⁵ As Calabresi has pointed out, “what is ‘cheap’ and what is ‘costly’ itself derives from the tastes and values of society,”²⁹⁶ including the needs to foster technological innovation while protecting vulnerable individuals from avoidable harms.

In the context of defamation by hallucination, a reasonable prompter or user who has been properly warned about the inevitability of LLM hallucinations (and the dangers of uncritically accepting AI outputs) may *in some instances* be the “cheapest cost avoider,” not to mention the most culpable actor in causing reputational harm to any persons defamed by the hallucination. Although the developer or deployer may have more control over the initial design of the LLM,²⁹⁷ LLM users should be aware of the potential for defamatory output, suspend credulity, and verify all outputs through readily available alternative sources before “republishing” (or otherwise disseminating) information that might harm the reputation of a third party. Failure to do so is at least negligent, and the negligence of the human user arguably supersedes any culpable conduct of the AI designers, trainers, or the chatbot itself—at least when users have been properly warned.

The same is true of users who act upon defamatory hallucinations by discriminating against the victim, whether as an employee, business partner, or social relation.²⁹⁸ The law does not generally recognize a cause of action against those who

²⁹⁵ See generally CALABRESI, *supra* note 139, at 1–340; Ronald Coase, *The Problem of Social Cost*, 3 J.L. & ECON. 1 (1960).

²⁹⁶ Calabresi & Smith, *supra* note 139, at 185.

²⁹⁷ See Sharkey, *supra* note 94 (contending that in many instances, particularly those involving physical injuries caused by AI products, the AI developer will be “the party best poised to avoid and/or mitigate accidents”).

²⁹⁸ For further discussion of the problems posed by those who simply act upon defamatory hallucinations by discriminating against the person defamed, see *supra* note 145 and accompanying text. Unfortunately, there will be those who simply act on defamatory hallucinations. An employer, for example, may refuse to hire a job applicant based on a hallucinated response about them. Such cases already exist: Employers sometimes rely on false gossip or inaccurate Internet search results to reject applicants. Such cases mostly go undetected. If they are indeed detected, the rejected applicant may be able to sue for discrimination based on a protected characteristic or against protected speech, if there is evidence that (for instance) the employer selectively targeted the applicant for

base decisions on faulty information; however, in some instances, such unjustified reliance may lead to conduct or actions that constitute defamation; in other instances, especially in employment contexts, unjustified reliance on false information may result in unlawful discrimination.²⁹⁹ Regardless, the problem of people acting upon unverified rumors is not new,³⁰⁰ and it is not clear that imposing liability on the AI system will appreciably ameliorate or deter such behavior.

IV. APPLYING LIABILITY RULES TO REASONING MODEL DEFAMATION

The preceding Part identified policies that defamation law should strive to achieve in adapting existing doctrines and principles to the problem of defamation

closer scrutiny on those bases. For example, Title VII of the Civil Rights Act forbids hiring discrimination based on sex, color, national origin, race, and religious practice. Discrimination based on marital status, sexual orientation, age, or disability status is often unlawful as well. *See, e.g.*, Jill Barth, *Federal Court to Consider AI in Hiring as Workday Bias Case Advances*, HR EXEC. (June 3, 2025), <https://perma.cc/G8JG-RLHJ> (describing a class action lawsuit brought against a company providing an AI hiring tool that allegedly discriminated against job applicants based on race, age, and disability). Concerns about bias in algorithmic decision-making tools have led to the passage of new employment laws and ordinances. For example, a New York City ordinance requires employers to disclose their use of automated employment decision tools, including those using artificial intelligence; employers must also conduct annual bias audits on these decision-making tools. *See* N.Y.C., N.Y., LOCAL LAW 144 (2021). A recently passed California law similarly requires transparency and human oversight for the use of AI hiring tools. *See* CAL. LAB. CODE §§ 1520–1539 (2025). Such regulations may be necessary to deter employers from engaging in AI-enabled discrimination based on unverified information about candidates that falls within a protected category. Even if AI systems were subjected to liability for failure to remove defamatory hallucinations after notice, such liability would only modestly ameliorate the harms caused by those who are unreasonably discriminating based on hallucinated content.

²⁹⁹ A few states have laws that prevent employers from firing employees for engaging in lawful activity while they are not on the job. *See, e.g.*, COLO. REV. STAT. § 24-34-402.5 (2024) (providing a cause of action to an employee terminated for lawful off-duty activities unless the employer’s restriction on such activities was related to a bona fide occupational requirement or “reasonably and rationally related to the employment activities and responsibilities of a particular employee or particular group of employees”; or the restriction is “necessary to avoid a conflict of interest” or “the appearance of such a conflict”).

³⁰⁰ For example, a dating app that was released in 2023 allows women to share information with each other about the men they date. Some of the men “reviewed” on this app, called Tea (as in “spill the tea”), alleged that the app’s contents defamed them or invaded their privacy. Isabella Kwai, *What to Know About the Hack at Tea, an App Where Women Share Red Flags About Men*, N.Y. TIMES (July 26, 2025). But men said the same thing of a 2005 website called DontDateHimGirl.com. *See* Lizette Alvarez, *(Name Here) Is a Liar and a Cheat*, N.Y. TIMES (Feb. 16, 2006).

by hallucination, particularly as the problem manifests in the outputs of LLM reasoning models. Of course, not all policy goals can be maximized simultaneously, and the application of liability rules must reflect a balance among competing objectives. In this Part, we evaluate the two most prominent proposals emerging from “first-generation” defamation-by-hallucination scholarship. After discussing their merits and limitations, we propose additional alternatives.

A. *The Limits of Treating Hallucinations as Negligent-Design Defects*

Legal scholar Nina Brown and others³⁰¹ have suggested that a negligently designed chatbot or LLM that produces defamatory hallucinations may be analogized to other negligently designed products,³⁰² such as, perhaps, cars. The appeal of the analogy between a hallucinating LLM and a defectively designed product is obvious: As Brown argues, imposing negligence-based liability would force AI producers to internalize the costs of the reputational harms they inflict, just as product suppliers must internalize the costs of physical harms their products cause. This approach also sidesteps the problem of how to apply scienter to a non-human bot, much as the judge did in *Walters*. Brown concedes that imposing strict liability—a standard often used for defective products—would be out of step with modern defamation principles,³⁰³ a proposition we explore below.

³⁰¹ Volokh, *supra* note 26, at 489, 523–24 (“[T]he question would be whether the company was negligent, and the answer would be analogous to the analysis of a negligent design product liability claim.”); Peter Henderson, Tatsunori Hashimoto & Mark A. Lemley, *Where’s the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589, 649 (2023) (“Instead, we think an AI’s liability should be judged objectively, not subjectively. An AI should be liable for false speech only if it was not designed using standard practices intended to mitigate that risk.”). *See also* Bambauer, *supra* note 45, at 349–58 (analyzing existing duty rules for the AI context, separating hypotheticals where (1) the AI gives misinformation causing harm to the user; (2) the AI gives misinformation causing harm to a third party (via the user’s conduct); and (3) the user does not use AI when it would have averted physical harm by providing accurate information).

³⁰² Brown, *supra* note 45, at 409 (“[T]he 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.”).

³⁰³ *Id.*; *see also id.* at 406 (“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.”); Volokh, *supra* note 26, at 489, 525 (arguing design defect product liability law should be at least applicable by analogy here, even if the fault standard would need to be shifted to align with libel law’s disfavor of strict liability).

While superficially appealing, analogizing LLM hallucinations to negligent design defects fails to account for the unique features of information products and the nature of generative AI at its current stage of development. Under Brown's approach, a plaintiff harmed by a defamatory hallucination would be able to recover damages by showing that an AI "model was designed in a way that made it likely to generate defamatory statements."³⁰⁴ The problem with this argument, which treats hallucinations as defects, is that all models currently generate defamatory statements.

Even if the law were to require plaintiffs to show that a model was *unreasonably* likely to generate defamatory statements, it is not clear how that would be evaluated. Reasonableness is relative. Would the mere fact that one model generates hallucinations more often than another mean that the first was unreasonably designed? Would it matter what type of hallucinations were most prevalent in a given model? In evaluating the design of autonomous vehicles, decision-makers are inevitably comparing their operation to that of non-autonomous vehicles rather than simply to other autonomous ones, but in the case of LLMs, we may not yet have enough understanding, even with expert help, to know what is reasonable.

In fact, one might even contend that today's AI models are "unavoidably unsafe,"³⁰⁵ at least at this stage in their development. Though the notion of the unavoidably unsafe product was developed in the context of products posing potential risks of physical harm, not reputational, the idea is the same: Makers of certain inherently risky but socially useful products should not be subject to liability for failing to eliminate their inherent risks. A perfectly safe steak knife, for example, is a contradiction in terms.

Moreover, tort law has long treated the provision of information products differently than the provision of products that cause physical harm, and for good reasons: Speech produces intangible and broad-based benefits that justify privileging the negligent information provider over other product producers.³⁰⁶ Some AI

³⁰⁴ Brown, *supra* note 45, at 410. Brown contends that negligence law is sufficiently adaptable to handle claims based on a failure to use "reasonable care in programming, training, and deploying the chatbot." *Id.* at 422. A plaintiff might prove a design flaw by showing, for example, that "programmers prioritized generating sensational or controversial content over accurate and non-biased content." *Id.* at 410–11.

³⁰⁵ See RESTATEMENT (SECOND) OF TORTS § 402A cmt. k (1986).

³⁰⁶ See *supra* notes 102–106 and accompanying text for further elucidation of this argument.

companies are among the wealthiest businesses in the world. Even assuming these companies could internalize all of the costs they impose on individuals and society, the law might still choose, for public policy reasons, not to force them to do so. For example, making AI producers strictly liable for defamatory hallucinations might encourage unduly broad censorship of AI outputs, restricting public access to information. And imposing strict liability might also encourage AI developers to release only models that minimize hallucinations, at the expense of developing models that maximize creative problem-solving capability.

Furthermore, a hallucinating chatbot is not necessarily a negligently designed or deployed chatbot. Applying risk-utility analysis of chatbot design or operation on a case-by-case basis seems unworkable, or at least unwise, because even if some LLMs are less prone to generate hallucinations, it does not mean that the ones that are more prone are defective.³⁰⁷ In some contexts, greater tolerance for hallucinations may be a design tradeoff deliberately made by the model's developers in order for it to generate outputs that are more imaginative, nuanced, or responsive to novel tasks.³⁰⁸ Or designers may reasonably choose to give users a choice about how much hallucination to tolerate.

A critic might argue that courts and juries can and should weigh the multiple risks and potential utility of various LLMs in order to push AI designers toward "safer" designs.³⁰⁹ But generative AI is still in such a nascent state, and is developing

³⁰⁷ RESTATEMENT (THIRD) OF TORTS: PRODS. LIAB. § 2(b) (1998) ("A product is defective when, at the time of sale or distribution, . . . the foreseeable risks of harm posed by the product could have been reduced or avoided by the adoption of a reasonable alternative design . . . and the omission of the alternative design renders the product not reasonably safe.").

³⁰⁸ See *supra* Part II.B. See ANTHROPIC, *supra* note 255, at 15 ("For CoT monitoring to be most effective, the CoT must be a faithful and complete reflection of the way the model reached its conclusion and generated a user-facing response. This means that the model's CoT must highlight the key factors and steps behind its reasoning. If CoT is not fully faithful, then we cannot depend on our ability to monitor CoT in order to detect misaligned behaviors, because there may be important factors affecting model behavior that have not been explicitly verbalized.").

³⁰⁹ For examples of the difficulty in determining what constitutes a reasonable design (or a successful engine), see Lisa Dunlap et al., *VibeCheck: Discover & Quantify Qualitative Differences in Large Language Models* (Apr. 19, 2025), <https://perma.cc/5QPX-QZ8P>; see also WSJ Staff, *AI Buzzwords You Need to Know Now*, WALL ST. J. (Feb. 20, 2025) ("Tech industry insiders are increasingly relying on intuition, rather than hard data, to judge which AI chatbots are best."). See generally *Indianapolis Colts, Inc. v. Metro. Baltimore Football Club Ltd.*, 34 F.3d 410, 415 (7th Cir. 1994) (describing the battle of the experts in trademark law as "frequently unedifying").

so rapidly, that it is hard to imagine how effectively judges and juries could weigh risks and utility—even with expert guidance.³¹⁰ At present, the pace of technological advancement in AI is so rapid that by the time an alleged “negligent design” case reaches a jury, the outcome would hinge on design choices and tradeoffs made concerning obsolete LLM models and outdated technical standards. Further, applying the concept of negligent design invites legal decision-makers to treat LLMs as static algorithmic products producing predictable outcomes, perhaps ignoring the unpredictable “black box” from which inevitable errors sometimes emerge. At present, even technologists struggle to predict how frequently a model will hallucinate and how “serious” the hallucination will be, and even the “safest” models generate hallucinations. At least at this juncture, it seems fair to question the institutional competence of judges and juries to evaluate LLMs as information products.

Focusing unduly on LLMs’ design also ignores the important role of LLM users in contributing both to the production of defamatory outputs and the wider dissemination of those outputs. Liability rules should further incentivize the rapidly increasing sophistication of LLM users, rather than treating them analogously to hapless consumers injured by harmful products.

In cases of defamation by hallucination, the person injured is not the consumer or user of the defective product but is instead a bystander. The active role played by the user in contributing to the bystander’s harm argues, at a minimum, for sharing responsibility for that harm between the LLM developer/provider and the LLM user. Legal decision-makers should apply liability rules in cases concerning AI in a manner that encourages users to verify LLM outputs before spreading them to other audiences. Regardless, treating hallucinations like design defects would place undue faith in the ability of judges and juries to keep pace with the newest design tweaks.³¹¹

³¹⁰ Dunlap et al., *supra* note 309; WSJ Staff, *supra* note 309; *Indianapolis Colts*, 34 F.3d at 415.

³¹¹ As we argue at *supra* notes 295–300 and accompanying text, users play a significant role in contributing to the reputational harms stemming from defamatory hallucinations, and liability rules should encourage users to exercise reasonable care for preventing such harms. In addition to playing an independent role in protecting others from hallucinatory harms, users can (and do) apply pressure on AI developers to continue innovating toward safety, especially where current premium models charge month to month and involve little to no switching costs.

B. The Limits of Notice-and-Blocking Based Liability

Eugene Volokh has made one of the most compelling proposals for assessing the liability of AI model producers for defamation by hallucination. Volokh posits that when a company such as OpenAI receives notice that its LLM is making false assertions about an individual, or that the documents supplied by its model are fabricated and defamatory, the company should add code to block those particular defamatory allegations.³¹² In this frame, OpenAI's failure to block future hallucinations about the individual, if such blocking is feasible, would constitute actual malice and liability should follow. Other scholars have reached similar conclusions.³¹³

As detailed below, our proposal to treat LLM providers as distributors substantially aligns with Volokh's. Like Volokh, we worry about the prospect that takedown notices might be gamed by the rich and powerful to block public access to information. Also like Volokh, we are concerned that "blocking" might be over- or underinclusive. Yet we agree that it is worth tolerating a degree of over- or under-inclusivity in order to protect individuals from hallucination-caused harms.

We question, however, whether "blocking" of specific, targeted content is as feasible as Volokh seems to assume. Volokh contends that once an AI provider is notified that its model has made a false and defamatory factual assertion about a particular individual, "[p]resumably the company could then add code that would prevent these particular allegations—which it now knows to be false or at least likely false—from being output."³¹⁴ To be sure, OpenAI can block any searches about an individual with a "hard-coded name filter," and there is evidence that OpenAI has already done so in response to complaints.³¹⁵

³¹² Volokh, *supra* note 26, at 514–15.

³¹³ See Henderson, Hashimoto & Lemley, *supra* note 301, at 641 ("[A] company that is aware its software is regularly generating a particular false statement and does nothing about it may be liable.").

³¹⁴ Volokh, *supra* note 26, at 514–15.

³¹⁵ For example, when journalists for Ars Technica asked ChatGPT to respond to their searches concerning "Brian Hood," an Australian mayor who was probably the first person to sue OpenAI over a defamatory hallucination, ChatGPT responded: "'I'm unable to produce a response' or 'There was an error generating a response.'" ChatGPT would then end the chat session. Ars Technica discovered that ChatGPT provides similar responses if asked to comment about certain people,

This application of the “notice-and-blocking” approach is best understood as abstention.³¹⁶ At first glance, the trade-offs attached to this approach seem obvious: It is relatively easy for an LLM producer to simply block *any* answer to queries regarding particular names. When they do, they eliminate hallucinations at the cost of instituting complete censorship on those subjects. Therefore, notice-and-blocking may place a “kill switch” in the hands of the sophisticated and unscrupulous. For example, one teacher assigned his students a math problem. Unbeknownst to the students, the teacher had inserted an invisible, white-text watermark containing a name that he knew to be blocked by a hard-coded name filter. The watermark prevented his students from using ChatGPT to generate a response for their homework.³¹⁷ This strategic use of hard-coded name filters demonstrates the gamesmanship that may follow as a consequence of over-broad abstention. As “blocked names” became more pervasive (or as the public became more aware of these filters), additional similar applications of blocked names as a “kill-switch” would be likely to follow.³¹⁸

One response may be to ask, “What about targeted blocking?” Professor Volokh has suggested that an AI producer may add post-processing code “to prevent particular allegations from being output.”³¹⁹ One example of using a post-processing parameter to eliminate defamatory outputs is the logit-bias parameter. Rather than completely blocking LLM responses containing a given name, the logit-

presumably because they have complained to OpenAI after being the subjects of defamatory hallucinations. Benj Edwards, *Certain Names Make ChatGPT Grind to a Halt, and We Know Why*, ARS TECHNICA (Dec. 2, 2024), <https://perma.cc/LA6B-KDL8>.

³¹⁶ Abstention is the refusal of LLMs to provide an answer to a query. See Wen et al., *supra* note 230, at 11.

³¹⁷ Edwards, *supra* note 315. See also Riley Goodside (@goodside), X (Dec. 2, 2024, at 12:07 PM), <https://perma.cc/93YX-CXL5> (the teacher’s post).

³¹⁸ For those sympathetic to malicious compliance, this may not be a bad thing. As Benj Edwards notes, such hard-coded name filters may prevent ChatGPT from answering questions about articles containing a “blocked” name or processing a website with a blocked name added to its text. This could provide a sort of “opt-out” for authors uninterested in training future AI models to mimic their work. Edwards, *supra* note 315. See also Ashley Belanger, “*Torrenting from a Corporate Laptop Doesn’t Feel Right*”: *Meta Emails Unsealed*, ARS TECHNICA (Feb. 6, 2025), <https://perma.cc/2NJV-US73> (discussing a court filing alleging that Meta torrented a dataset containing tens of millions of copyrighted books in order to train its Llama model).

³¹⁹ Volokh, *supra* note 26, at 515.

bias parameter eliminates the occurrence of a particular name alongside a particular phrase. The logit-bias parameter works by modifying the likelihood of specified tokens appearing in a model generated output.³²⁰ For example, if a model erroneously says, “Person X embezzled funds from her company,” a logit-bias parameter might block future statements containing “Person X” and “embezzled.” As Volokh acknowledges, this type of filtering alone is unlikely to prevent defamatory hallucinations where LLMs output subtly different answers in response to the same query.³²¹ In this context, syntactic filters present a Hydra issue: Block one phrase, and the model may sprout two more with the same meaning.

Given the difficulty of implementing targeted blocking, one may wonder whether prompters should seek a private solution to avoid defamation liability. The notion of third party “post-processing filters” is already being tested. CaliberAI offers its services to journalists who are hoping to identify defamatory hallucinations before they are republished. The utility of these services appears to be limited. Neil Brady, the founder of CaliberAI, said “that while his company’s defamation filter is ‘unique,’ it isn’t fast and doesn’t catch all defamation because *fact-checking defamatory statements is not yet ‘really computationally possible.’*”³²² As Brady’s comment implies, CaliberAI can tell you whether a statement looks defamatory

³²⁰ See *Using Logit Bias to Alter Token Probability with the OpenAI API*, OPENAI (Aug. 2025), <https://perma.cc/9ZTQ-FZ6S>. But see @animesh, *Can We Block Certain Keywords in Output?*, OPENAI CMTY. F. (June 2021), <https://perma.cc/HKZ8-6PB3> (OpenAI community forum in which blocking a set of words is described as a “bandaid to prevent the model from talking about unwanted topics”).

³²¹ Volokh, *supra* note 26, at 515 (“Nonetheless, some such reasonably protective solution seems likely to be within the capability of modern language recognition systems, especially since a company would only have to take reasonable steps to block the regeneration of the material, not perfect steps.”).

³²² Ashley Belanger, *Will ChatGPT’s Hallucinations Be Allowed to Ruin Your Life?*, ARS TECHNICA (Oct. 23, 2023), <https://perma.cc/HX26-H6HW> (emphasis added). For what it’s worth, CaliberAI, a tool built to identify defamatory hallucinations, provides an advisory very similar to a standard AI disclaimer: “At CaliberAI we aim to assist publishers by helping to minimise risk of publication of defamatory or harmful content. When our systems classify phrases or statements, judgements are being made about linguistic structures, and not subject matter. Our technology is built neither to censor nor allow for unrestricted free speech, but to strike a balance by augmenting human editing skills. Ultimately, the decision to publish rests with the user.” *Advisory*, CALIBERAI, <https://perma.cc/C5TB-FQGH>.

(whether its syntax reflects that of a standard defamatory statement), but it cannot tell you whether the semantic meaning of the statement is false.

Consider the simple formula underlying CaliberAI's design:³²³ CaliberAI coded its product with many examples of what it considered to be defamatory statements (although some of these appeared to be non-factual opinion). The product then "flags" negative, qualitative statements produced by LLM outputs, but it does not verify the factual content of such outputs. In statements to the media, Brady claimed his product could have detected one of the sentences from the hallucinated complaint that had triggered the *Walters* lawsuit³²⁴—which is really to say that it could have identified that the *Walters* statement contained a subjective clause, "Mark Walters is believed to have," and a taboo phrase, "misappropriated funds." At its current stage of development, CaliberAI represents a bespoke Grammarly more than it does a comprehensive screening tool for defamatory outputs.

Two years ago, Eugene Volokh expressed faith that a "reasonably protective solution" was "within the capability of modern language recognition systems."³²⁵ Reasoning models represent a significant step toward that solution. Volokh wondered whether a company "can show that (1) it can design a system that can perform at nearly the 90th percentile on the bar exam, but that (2) checking the system's output to see if it includes a particular person's name in an assertion about an embezzlement conviction is beyond the company's powers."³²⁶ The answer, surprisingly, may be that even though reasoning models are making progress in

³²³ CaliberAI described the process as follows: "Vogel started with a simple formula, generating a series of defamatory statements that combined a subjective clause with a taboo phrase. Brady gave an example of a simple statement that would be flagged as defamatory: 'Everybody knows that Joe Biden is corrupt.' From there, Vogel created what Brady called a 'gold standard data set,' then CaliberAI hired a team of annotators to add 100,000 annotations and emulate the data so that the defamation filter would catch more content." *Frequently Asked Questions: How Do We Decide What Is Defamatory/Harmful?*, CALIBERAI, <https://perma.cc/VH2E-ANC4>.

³²⁴ "Brady said that CaliberAI would have flagged ChatGPT's output falsely claiming that 'Mark Walters is believed to have misappropriated funds from a gun Second Amendment Foundation in excess of \$5,000,000' as defamatory." Belanger, *supra* note 322.

³²⁵ Volokh, *supra* note 26, at 516.

³²⁶ *Id.*

“checking system outputs” for errors, certain types of errors are, indeed, still beyond the power of the AI company to eradicate.³²⁷

The best approach to addressing these issues appears to be training the model’s policy compliance into the generation of the output itself. This approach is explicitly deployed in OpenAI’s latest reasoning models.³²⁸ Deliberative alignment is a training paradigm that “teaches LLMs to explicitly reason through safety specifications before producing an answer.”³²⁹ Moreover, reasoning models utilizing these safeguards are now available to free users in ChatGPT and appear poised to become more broadly available for little to no cost in the immediate future.³³⁰ DeepSeek-R1 is another example of a reasoning model available to the public at no cost.³³¹

C. *The Case for Imposing Statutory Duties on AI Producers*

Despite progress toward creating safer models, there is a strong argument for imposing a statutory duty on AI producers to warn users regarding the inevitability and perhaps the frequency of hallucinations. Even in the most “accurate-seeming” models,³³² hallucinations will occur, and users should verify outputs to mitigate potential harm. Yet, for the LLM producer, market incentives cut against providing such warnings. In fact, AI producers often tout the accuracy of LLMs in an effort to gain a competitive advantage over other producers.³³³ Users want more accurate models and LLM producers have an incentive to falsely portray their LLMs as

³²⁷ See *supra* Part II.

³²⁸ See OPENAI, *supra* note 283. See also Melody Y. Guan et al., *Deliberative Alignment: Reasoning Enables Safer Language Models* (Jan. 8, 2025), <https://perma.cc/D7NS-ULNE>.

³²⁹ Guan et al., *supra* note 328, at 1.

³³⁰ *OpenAI o3-Mini*, OPENAI (Jan. 31, 2025), <https://perma.cc/E6QR-D3X2> (“Starting today, free plan users can also try OpenAI o3-mini by selecting ‘Reason’ in the message composer or by regenerating a response. This marks the first time a reasoning model has been made available to free users in ChatGPT.”).

³³¹ See CHATBOT.APP, *supra* note 208.

³³² See Magesh et al., *supra* note 24, at 2 n.2 (citing to official statements from Lexis, Casetext, and Thomson Reuters, each claiming to offer a hallucination-free AI-based legal research tool). After recent tests found that OpenAI’s o3 and o4-mini models produced more hallucinations than its o1model, a company spokeswoman promised that the company would “continue our research on hallucinations across all models to improve accuracy and reliability.” Metz & Weise, *supra* note 214 (quoting Gaby Raila of OpenAI).

³³³ Magesh et al., *supra* note 24, at 2 n.2.

“safe.” Thus, it makes sense to find a means of requiring producers to provide feasible, general warnings to all users about the existence and prevalence of harmful hallucinations.

However, tort law is not the easiest and most effective way to impose a duty to warn all LLM users about the existence and prevalence of hallucinations and their potential for harm. The straighter route is undoubtedly to implement statutory laws that would mandate transparency. A statutory duty could be tied to a corresponding immunity from liability, but even without such a tie, the creation of a statutory “public duty” of educating potential users about AI’s dangers and limits would be valuable in mitigating potential reputational harms.

Although somewhat beyond the scope of this article, lawmakers could also supplement user transparency laws with new laws and regulations requiring transparency about design choices, training data, and testing protocols—all in the interest of mitigating potential harms stemming from the inevitable errors of LLMs.³³⁴

It would also be wise for lawmakers to require LLM developers and producers to retain search data for at least a couple of years, which should be long enough for defamation victims to discover the existence of a defamatory output concerning them (if they are going to).³³⁵ The retention of search records would help such victims prove harm, if any exists.

An important caveat to this argument is that data retention obligations come with privacy and data security tradeoffs. The longer sensitive data is retained, the more vulnerable it is to being breached, misused, or exposed. In other words, data-retention obligations create privacy and data-security concerns. Consider a hypothetical “high-privacy LLM,” similar to DuckDuckGo’s privacy-focused search engine that automatically deletes all its user prompts and logs to ensure user anonymity. Such a high-privacy LLM, while providing a clear benefit to privacy-conscious users, would be in direct violation of a law that mandated data retention.

³³⁴ Providing transparency to users may help them understand the dangers of unthinking reliance on AI outputs. It is worthwhile distinguishing between developer-side CoT transparency and user-side CoT transparency. The former may assist further innovations for accuracy more than the latter. But user-side transparency may help train users about the dangers of relying on LLM outputs without verification. See Nipuna Thalpage, *Unlocking the Black Box: Explainable Artificial Intelligence (XAI) for Trust and Transparency in AI Systems*, 4 J. DIGIT. ART & HUMANS. 31 (2023).

³³⁵ See OPENAI, *supra* note 271 (data-retention policy).

D. LLM Producers as Quasi-Distributors—and the Special Instance of Chains of Thought Hallucinations

Shifting from statutory to common law, defamation law should be calibrated to address the issue of defamatory hallucinations—particularly those that arise in chain-of-thought reasoning. Our primary recommendation is that the common law apply the publisher-distributor distinction with sensitivity to how LLMs actually operate.

As we argued in our discussion of the *Walters* case, LLM producers should not be treated as publishers of a defamatory factual assertion when their models respond with a defamatory hallucination in response to a user query. Unlike true publishers, AI producers do not themselves consciously select and create original content but instead provide a technological tool that creates an algorithmic “remix” of information gleaned from training data or external sources. As my colleague Jane Bambauer has explained, this makes LLM outputs more like the products of search engines than the products of newspapers. Though LLM producers can sometimes control in advance user access to certain topics—such as by blocking, say, references to Tiananmen Square³³⁶—they currently lack the degree of editorial control exercised by those independent content producers the law treats as “publishers.” Accordingly, defamation law should not impose on LLM producers a duty to confirm the truth of what their models produce, at least not prior to “publication.”

Instead, those applying defamation law’s categories to LLMs should recognize the vital role of LLMs as content-producing tools, and the limited ability of LLM producers to exercise day-to-day, advance control over their model’s responses to the millions, or eventually billions, of user prompts it receives. Indeed, LLM producers often lack the technical capacity to exercise tailored editorial control to prevent further hallucinations even *after* they learn of the occurrence of such errors. Regardless, at the current stage of LLM development, the provision of an LLM for public use cannot be seen as an implicit warranty of accuracy from its developer, and the public policy concerns listed in the last section support treating LLMs as distributors, or quasi-distributors.

³³⁶ See, e.g., James T. Areddy & Isabella Simonetti, *DeepSeek’s Chatbot Works Like Its U.S. Rivals—Until You Ask About Tiananmen*, WALL ST. J. (Apr. 6, 2025).

This approach reflects the insights gleaned from studying defamation law's application to the telegraph. Defamation law chose, for policy reasons, to treat telegraph operators as "passive conduits" of information, despite their capacity to exercise control over senders' messages. Just as imposing liability on telegraph operators for failing to exercise editorial control over all messages they received would have hindered the development and use of that revolutionary technology, so too would treating LLM producers as publishers disrupt the development and use of CoT-based systems. At this stage in the innovation cycle, treating LLM producers as akin to distributors acknowledges their vital role in providing users with access to a new and powerful information-gathering technology; this categorization may also enable LLM producers to respond more innovatively to market demands for accuracy, safety, and transparency.

In applying the distributor rule, it is also worth remembering another lesson from the history of the telegraph. Defamation law refused to impose liability on telegraph operators who were "innocent disseminators," yet continued to impose liability on the culpable senders of defamatory telegraphs. By the same token, defamation law's similar refusal to impose liability for defamatory hallucinations on AI producers in all but rare instances would not mean that defamation victims will go uncompensated; the negligent LLM *user* who disseminates defamatory hallucinations without verifying them is still available to be sued.³³⁷

This approach has special saliency as applied to hallucinations appearing in "chains of thoughts" or "chains of verification" produced by LLMs. Even if courts decide as a general matter to treat AI producers as publishers, hallucinations confined to an LLM's internal reasoning process (for instance, false assertions that appear fleetingly in the model's intermediate chain of thought) should not be treated

³³⁷ Unfortunately, as the *LTL* case discussed at *supra* notes 141–144 and accompanying text demonstrates, there may still be LLM users who act on defamatory hallucinations, like by refusing to hire a job applicant based on a hallucinated response about them. Such cases already exist: Employers already rely on false gossip or inaccurate Internet search results to reject applicants, and such cases mostly go undetected. Regulation may be necessary to deter employers from engaging in AI-enabled discrimination against job candidates. See, e.g., Barth, *supra* note 298 (discussing various laws, such as Title VII of the Civil Rights Act, that forbid various types of employment discrimination). Imposing notice-and-takedown liability would only modestly ameliorate the harms from employers' unjustified reliance on hallucinating AI systems.

as publications and thus should not be actionable.³³⁸ Treating unfiltered, non-final reasoning as defamation would be doctrinally unsound and counter-productive. “Chain-of-verification” systems often generate and then discard erroneous content as a means of refining their final outputs. The LLM producer cannot feasibly monitor all chains of thought or chains of verification for signs of misalignment that might lead to defamatory hallucinations, but the “reasoning chains” help users understand model behavior, detect possible hallucinations, and avoid submitting prompts that produce them.

As CoT models become more ubiquitous, the public interest in CoT transparency grows stronger.³³⁹ Hence, imposing liability for a model’s transitory missteps

³³⁸ See *supra* Part II.A (illustrating how an LLM’s “chain of thought” may contain a defamatory hallucination that never reaches the user in the final output). In our example, the model’s intermediate reasoning falsely identified two professors as sexual harassers, only to exclude that error from its ultimate answer. Because those incorrect allegations remained within the AI’s internal verification process and were not communicated as facts to the prompter, they should not be deemed “published” for defamation purposes.

³³⁹ The argument made here is analogous to the argument for recognizing a “common interest” in sharing potentially defamatory information in order to further common goals of the speaker and listener. The prototypical common interest privilege shields the sharing of potentially defamatory information about a job candidate by a former employer with a prospective employer, but the common interest privilege protects the sharing of potentially defamatory information to serve broader public policy goals in a variety of circumstances. For example, some jurisdictions refuse to treat sharing of potentially defamatory information within a corporation as a publication, though this is a minority rule. See *Burks v. Pickwick Hotel*, 607 So. 2d 187, 189 (Ala. 1992) (reasoning that “[e]ssentially, such a situation would be like an entity talking to itself rather than to a third party”); *Bell v. Rogers*, 698 So. 2d 749, 756 (La. Ct. App. 1997) (“[S]tatements between employees, made within the course and scope of their employment, are not statements communicated or publicized to third persons so as to constitute a publication.”); *Rice v. Hodapp*, 919 S.W.2d 240, 243 (Mo. 1996) (“[C]ommunications between officers of the same corporation in the due and regular course of the corporate business, or between different offices of the same corporation, are not publications to third persons”); *Magnolia Petroleum Co. v. Davidson*, 148 P.2d 468, 471 (Okla. 1944) (holding that an employee did not publish statements made to a supervisor as part of an investigation of co-worker); *Woods v. Helmi*, 758 S.W.2d 219, 223 (Tenn. Ct. App. 1988) (“[C]ommunication among agents of the same corporation made within the scope and course of their employment relative to duties performed for that corporation are not to be considered as statements communicated or publicized to third persons.”); *Prins v. Holland-N. Am. Mortg. Co.*, 181 P. 680, 680–81 (Wash. 1919) (“For a corporation, therefore, acting through one of its agents or representatives, to send a libelous communication to

would perversely discourage the very design features that improve final output accuracy. Treating internal CoTs as “unpublished” preserves breathing room for AI developers to iteratively finetune transparency and self-checking mechanisms without fear that every mistaken thought by a model will trigger a lawsuit. It also hews to first principles: Further harm to reputation is avoided when no one else hears the falsehood. By digging into the CoT, a prompter assumes the responsibility of verifying its contents before reproducing them to a broader audience, and prompters who do reproduce them should be treated as if they originated them.

Applying the distributor approach advocated above, LLM producers should be responsible for harms arising from reputation-damaging hallucinations if they are notified or otherwise become aware that their LLM has produced such a hallucination about a particular individual and are capable of preventing the LLM from producing the same defamatory hallucinations in response to subsequent user prompts. In other words, if blocking of specific, defamatory falsehoods is feasible, an LLM producer should prevent its model from repeating those falsehoods.

Currently, however, it is far from clear whether it is possible to remove a specific defamatory allegation,³⁴⁰ and just as defamation law does not penalize libraries that fail to remove entire books containing a single defamatory sentence from their shelves, we should not create rules that incentivize LLM producers to block entire topics or refuse to provide any information whatsoever about particular individuals. Distributor liability rules should be applied to protect the emerging role that AI reasoning models play in providing broad public access to information created by others. That role would be compromised if the models’ producers were incentivized to remove broad swaths of content upon notice that a single defamatory hallucination had occurred.

Treating LLMs as distributors will not deprive defamed plaintiffs of recourse. In many cases, however, it will mean that plaintiffs must sue those who negligently

another of its agents or representatives, cannot be a publication of the libel on the part of the corporation. It is but communicating with itself.”). The rationale for the minority rule on intracorporate communications is that no publication has occurred because the corporation is simply “communicating with itself” via its human agents. An analogous argument could be made in the CoT reasoning model context: Hallucinations within chains of thought should not be treated as publications because the hallucinations represent the reasoning model “communicating with itself” in preparation for producing its final output.

³⁴⁰ See *supra* Part IV.B for feasibility of blocking proposals.

spread AI-generated falsehoods about them. A user who has been warned about the propensity of LLMs to hallucinate but nonetheless prompts and then spreads LLM outputs without verification is akin to a journalist who publishes falsehoods supplied by an unreliable source.³⁴¹ Depending on the circumstances, such republication is not just negligent but manifests reckless disregard for the truth of the defamatory statement.³⁴² Imposing liability on users who negligently, recklessly, or knowingly disseminate defamatory hallucinations will have the salutary consequence of incentivizing all users to exercise reasonable care in using this new technology.

E. Dangers of Imposing Strict Liability

The legal system sometimes opts to apply strict liability to those whose activities or businesses inevitably produce harm. Such schemes often apply to activities that create statistically inevitable harms but provide net social benefit. Legal policymakers sometimes offer such producers a quid pro quo: To avoid the unpredictability and transaction costs associated with the tort system, the producers must compensate the individuals they harm, regardless of whether such individuals can prove fault. Such “no-fault” recoveries, however, are limited and relatively predictable. A prototypical example is workers’ compensation, in which workers can recover compensation for on-the-job injuries without proving the fault of their employers but their compensation is limited to more modest and predictable sums than a negligence-based tort regime would provide.

If AI hallucinations are inevitable and cannot be eradicated but continued development of AI technologies is of benefit to society, one might propose that lawmakers should force AI producers to internalize the costs of defamatory hallucinations by paying damages to those defamed, regardless of whether the AI producer was at fault for the reputational harms incurred. Damages could be capped at modest amounts, calculated in part based on how many LLM users received the defamatory results. Such a scheme would protect the victims of “bots gone bad” without imposing ruinous liability on bot creators.

Yet imposing a strict-liability statutory damages scheme for AI-generated defamation would invite unintended consequences. Given the inevitability of hallucinations and the low-to-nonexistent cost and effort their production requires, one can easily imagine users seeking, in bad faith, to systematically prompt chatbots to

³⁴¹ See, e.g., *St. Amant v. Thompson*, 390 U.S. 727 (1968).

³⁴² *Id.* at 730–33.

produce defamatory hallucinations in order to capitalize on a no-fault recovery system. A strategic actor could prompt an LLM with the names of a virtually unlimited pool of private individuals—automatically and *ad infinitum*—until a defamatory hallucination emerged. The actor would then contact the ostensibly defamed individual and offer to pursue a claim on their behalf—at least so long as the “victim” shared with the actor the proceeds of the claim. Prompters working to generate such claims might even frame their behavior as socially beneficial. But the unleashing of defamation claims through such a system would do little to serve defamation law’s traditional purposes and much to undermine innovation within and among AI systems.³⁴³

Some critics will fault this proposal for granting significant liability protection to a powerful industry that is extractive, exploitative, and insufficiently sensitive to the risks its technology imposes not just to vulnerable victims but to human flourishing more generally. After all, companies like OpenAI are getting rich by hoovering up copyrighted works without authorization.³⁴⁴ They threaten environmental destruction based on their rapacious energy consumption.³⁴⁵ Already, vulnerable users, including children, have died after becoming psychologically dependent on

³⁴³ There is also a resource cost to unnecessary AI searches. See Pranshu Verma & Shelly Tan, *A Bottle of Water Per Email: The Hidden Environmental Costs of Using AI Chatbots*, WASH. POST (Sep. 18, 2024).

³⁴⁴ Ian Stark, *Meta Gets Partial Win in AI-Teaching Copyright Case*, UPI (June 26, 2025), <https://perma.cc/59KT-WZP5>; Matt O’Brien, *Anthropic Wins Ruling on AI Training in Copyright Lawsuit but Must Face Trial on Pirated Books*, AP NEWS (June 24, 2025), <https://perma.cc/Y884-2T9T>; Gavin Doyle, *Universal and Disney Team Up to Fight Shocking Content Theft by New AI Tool*, MSN (June 11, 2025), <https://perma.cc/8J5S-8J54>; Bobby Allyn, *Judge Allows ‘New York Times’ Copyright Case Against OpenAI to Go Forward*, NPR (Mar. 26, 2025), <https://perma.cc/48YY-TFSN>.

³⁴⁵ Kate Crawford, *Generative AI’s Environmental Costs Are Soaring—and Mostly Secret*, NATURE (Feb. 20, 2024), <https://perma.cc/DRR6-F2VC>.

chatbots.³⁴⁶ Moreover, a recent study reported that daily use of AI to perform writing tasks saps ordinary users of critical thinking skills.³⁴⁷ At the same time, AI companies tout the revolutionary potential of LLMs while delivering flawed models that have the capacity to harm society in myriad ways, including by swamping the current information ecosystem with yet more misinformation.

Of these concerns, however, only the last even arguably comes within defamation law's ambit. Defamation law exists to protect reputations by vindicating those who suffer reputational injury in the form of dignitary, relational, and economic harms.³⁴⁸ While defamation law plays a role in protecting the information ecosystem from falsehoods that harm individuals, it does nothing to combat falsehoods writ large. It is simply not adequate to the task.

In reaching this conclusion, we accept that, for many purposes, scholars Margot Kaminski and Meg Leta Jones are correct in their assessment: AI should be understood "not as a speaker, nor as [a] generator of speech at scale, but as a risky complex system."³⁴⁹ As such, AI systems as a whole call for "powerful forms of intervention," mostly in the form of direct regulation. These regulations, such as "mandated design specs or more general recording and reporting requirements," should force "developers and users" to "lower[] risks to the population at large."³⁵⁰ Unlike Kaminski and Jones, we offer not a comprehensive approach to AI regulation or the harms caused by hallucinated content but a limited solution to a limited problem, a solution rooted in a deep understanding of defamation law's flaws and limitations.

³⁴⁶ SUSAN ALEGRE, HUMAN RIGHTS, ROBOT WRONGS 39 (2025) (describing case involving Belgian father in his early 30s who took his life after developing an intense relationship with a chatbot); Will McCurdy, *Meta, Character.AI Therapy Bots Under Fire for 'Endangering the Public'*, PCMag (June 15, 2025), <https://perma.cc/425R-RLDA> (describing case involving Florida teen who took his life after developing an intense relationship with a chatbot).

³⁴⁷ Andrew R. Chow, *ChatGPT May Be Eroding Critical Thinking Skills, According to a New MIT Study*, TIME (June 23, 2025), <https://perma.cc/7E5M-DXY8>.

³⁴⁸ Lidsky, *supra* note 106, at 92.

³⁴⁹ Margot E. Kaminski & Meg Leta Jones, *Constructing AI Speech*, 133 YALE L.J.F. 1212, 1248 (2024) (emphasis omitted).

³⁵⁰ *Id.*

CONCLUSION

Over the last millennium, defamation law has adapted to many new information technologies, including the printing press, the telegraph, the radio, the television, and the Internet. To do so, the common law often has adapted its rules to ensure that those who provide public access to the newest communications tools do not suffer crippling liability.

In this article we have attempted to unite the lessons of history with cutting-edge computer-science research in deciding what to do about the inevitable yet essential and sometimes valuable errors produced by AI reasoning models. The law must impose on LLM producers a duty to warn (and educate) their users about these inevitable errors and a duty to keep search records for a limited time. But once AI producers comply with these obligations, likely imposed by statute, the common law should not treat models like ChatGPT as the “publisher” of the hallucinatory communication. This is true whether hallucinations emerge in an LLM’s chain of thought, chain of verification, or even its final output. Instead, taking inspiration from the creativity of common-law judges in adapting tort doctrines to technologies such as the telegraph, judges today should treat LLM producers as information distributors—at least at this juncture. Meanwhile, defamation law can still safeguard reputation by treating LLM users who intentionally or negligently spread hallucinated and defamatory falsehoods in much the same way the law currently treats incompetent or unscrupulous journalists passing along information gleaned from unreliable sources.³⁵¹

³⁵¹ Special thanks to media law scholar Lili Levi of the University of Miami for raising this point in a discussion with one of the authors of this article.

