

**IN THE SUPERIOR COURT OF GWINNETT COUNTY
STATE OF GEORGIA**

MARK WALTERS,

Plaintiff,

v.

OPENAI, L.L.C.,

Defendant.

CIVIL ACTION NO. 23-A-04860-2

**BRIEF OF *AMICUS CURIAE* TECHNOLOGY LAW AND POLICY CLINIC
AT NEW YORK UNIVERSITY SCHOOL OF LAW
IN SUPPORT OF NEITHER PARTY**

Clare R. Norins*
Georgia Bar No. 575364
FIRST AMENDMENT CLINIC
University of Georgia School of Law
P.O. Box 388
Athens, Georgia 30603
(706) 542-1419
cnorins@uga.edu

Counsel for Amicus Curiae

**Local counsel for Technology Law and Policy
Clinic authors Micah Musser, Catherine
Wang, and Jake Karr*

December 12, 2024

TABLE OF CONTENTS

TABLE OF AUTHORITIES ii

INTEREST OF *AMICUS CURIAE*..... 1

SUMMARY OF THE ARGUMENT 1

ARGUMENT 2

 I. Defamation law should protect against reputational harms caused by generative AI outputs..... 2

 A. Powerful technology companies are increasingly embedding generative AI tools throughout the internet, creating risks of real and widespread harms. 3

 B. Reasonable users could believe generative AI outputs constitute statements of fact. .. 8

 II. Defamation plaintiffs should be required to allege specific unreasonably risky conduct taken by generative AI companies..... 11

 A. Imposing a fault requirement protects both generative AI companies’ and the public’s First Amendment interests in generative AI tools. 12

 B. Evaluating the fault of generative AI companies requires creatively borrowing from other bodies of law..... 13

 1. Products liability law can be used to evaluate the negligence of generative AI companies with regard to pre-deployment design decisions. 15

 2. Agency law can be used to evaluate the negligence of generative AI companies with regard to post-deployment oversight decisions. 18

 3. Both products liability and agency law can be used to evaluate particularly egregious conduct that may raise an inference of actual malice..... 21

CONCLUSION..... 24

TABLE OF AUTHORITIES

CASES

Banks v. ICI Americas, Inc.,
264 Ga. 732 (1994) 16, 18

Bd. of Educ. v. Pico,
457 U.S. 853 (1982)..... 13

Bollea v. World Championship Wrestling, Inc.,
271 Ga. App. 555 (2005) 8

Cianci v. New Times Publ’g Co.,
639 F.2d 54 (2d Cir. 1980)..... 9

Citizens United v. Fed. Election Comm’n,
558 U.S. 310 (2010)..... 12

Curtis Publ’g Co. v. Butts,
388 U.S. 130 (1967)..... 17

DaimlerChrysler Motors Co. v. Clemente,
294 Ga. App. 38 (2008) 18, 20

First Nat’l Bank of Boston v. Bellotti,
435 U.S. 765 (1978)..... 13

Gast v. Brittain,
277 Ga. 340 (2003) 8

Gertz v. Robert Welch, Inc.,
418 U.S. 323 (1974)..... 12

Gettner v. Fitzgerald,
297 Ga. App. 258 (2009) 18

Harcrow v. Struhar,
236 Ga. App. 403 (1999) 8

Harte-Hanks Communications, Inc. v. Connaughton,
491 U.S. 657 (1989)..... 23

Hosp. Auth. of Houston Cnty. v. Bohannon,
272 Ga. App. 96 (2005). 5

Hustler Mag., Inc. v. Falwell,
485 U.S. 46 (1988)..... 8

<i>Jones v. NordicTrack, Inc.</i> , 274 Ga. 115 (2001)	16
<i>Milkovich v. Lorain Journal Co.</i> , 497 U.S. 1 (1990).....	8, 9
<i>Moody v. NetChoice, LLC</i> , 144 S. Ct. 2383 (2024).....	12
<i>Mullinax v. Miller</i> , 242 Ga. App. 811 (2000)	20
<i>N.Y. Times Co. v. Sullivan</i> , 376 U.S. 254 (1964).....	21, 22
<i>St. Amant v. Thompson</i> , 390 U.S. 727 (1968).....	21
<i>Stanley v. Georgia</i> , 394 U.S. 557 (1969).....	12
<i>Universal City Studios, Inc. v. Corley</i> , 273 F.3d 429 (2d Cir. 2001).....	12

RESTATEMENTS

Restatement (Second) of Torts (Am. L. Inst. 1977)	17
Restatement (Third) of Agency (Am. L. Inst. 2006)	14, 18
Restatement (Third) of Torts: Prods. Liab. (Am. L. Inst. 1998).....	14, 16

OTHER AUTHORITIES

Aditi Bagchi, <i>Other People’s Contracts</i> , 32 Yale J. on Regul. 211 (2015).....	5
Amba Kak & Sarah Myers West, <i>ChatGPT and More: Large Scale AI Models Entrench Big Tech Power</i> , AI Now Inst. (Apr. 11, 2023), https://ainowinstitute.org/publication/large-scale-ai-models	4
Andrew Griffin, <i>What is QAnon? The Origins of Bizarre Conspiracy Theory Spreading Online</i> , Independent (Jan. 21, 2021), https://www.independent.co.uk/tech/what-is-qanon-b1790868.html	22
Anton Korinek & Jai Vipra, <i>Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT</i> (Ctr. on Regul. & Mkts. at Brookings, Working Paper No. 9, 2023), https://arxiv.org/pdf/2311.01550	3

Ashley Belanger, <i>Will ChatGPT’s Hallucinations Be Allowed to Ruin Your Life?</i> , Ars Technica (Oct. 23, 2023), https://arstechnica.com/tech-policy/2023/10/will-chatgpts-hallucinations-be-allowed-to-ruin-your-life	19, 23
Benj Edwards, <i>Certain Names Make ChatGPT Grind to a Halt, and We Know Why</i> , Ars Technica (Dec. 2, 2024), https://arstechnica.com/information-technology/2024/12/certain-names-make-chatgpt-grind-to-a-halt-and-we-know-why	19
Benj Edwards, <i>ChatGPT Hits 200 Million Active Weekly Users, but How Many Will Admit Using It?</i> , Ars Technica (Aug. 30, 2024), https://arstechnica.com/information-technology/2024/08/chatgpt-hits-200-million-active-weekly-users-but-how-many-will-admit-using-it	3
Celeste Kidd & Abeba Birhane, <i>How AI Can Distort Human Beliefs</i> , 380 Science 1222 (2023).....	7
David A. Hoffman, <i>Defeating the Empire of Forms</i> , 109 Va. L. Rev. 1367 (2023)	5
<i>Disrupting Deceptive Uses of AI by Covert Influence Operations</i> , OpenAI (May 30, 2024), https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations..	19
<i>Disrupting Malicious Uses of AI by State-Affiliated Threat Actors</i> , OpenAI (Feb. 14, 2024), https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors ..	14
Emily M. Bender et al., <i>On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?</i> , Proc. 2021 ACM Conf. on Fairness, Accountability & Transparency 610	7
Eugene Volokh, <i>Large Libel Models? Liability for AI Outputs</i> , 3 J. Free Speech L. 489 (2023).....	9, 10
Francesca Cabiddu et al., <i>Why Do Users Trust Algorithms? A Review and Conceptualization of Initial Trust and Trust over Time</i> , 40 Eur. Mgmt. J. 685 (2022).....	7
<i>GPT-4 Is OpenAI’s Most Advanced System, Producing Safer and More Useful Responses</i> , OpenAI, https://openai.com/index/gpt-4	9
Guido Zuccon et al., <i>ChatGPT Hallucinates when Attributing Answers</i> , Proc. Ann. Int’l ACM SIGIR Conf. on Rsch. & Dev. Info. Retrieval Asia Pac. Region 46 (2023).....	6
Helen Toner, <i>What Are Generative AI, Large Language Models, and Foundation Models?</i> , Ctr. for Sec. & Emerging Tech. (May 12, 2023), https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models	1
Helena Kudiabor, <i>AI Tool Helps People with Opposing Views Find Common Ground</i> , Nature (Oct. 17, 2024), https://www.nature.com/articles/d41586-024-03424-z	13

Junyi Li et al., <i>HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models</i> , Proc. 2023 Conf. on Empirical Methods Nat. Language Processing 6449.....	6
Katharina Buchholz, <i>The Extreme Cost of Training AI Models</i> , Forbes (Aug. 26, 2024), https://www.forbes.com/sites/katharinabuchholz/2024/08/23/the-extreme-cost-of-training-ai-models	3
<i>Learning to Reason with LLMs</i> , OpenAI (Sept. 12, 2024), https://openai.com/index/learning-to-reason-with-llms	17
Matt O’Brien, <i>Chatbots Sometimes Make Things Up. Is AI’s Hallucination Problem Fixable?</i> , Associated Press (Aug. 1, 2023), https://apnews.com/article/artificial-intelligence-hallucination-chatbots-chatgpt-falsehoods-ac4672c5b06e6f91050aa46ee731bcf4	7
Matt O’Brien, <i>Microsoft Bakes ChatGPT-like Tech into Search Engine Bing</i> , Associated Press (Feb. 7, 2023), https://apnews.com/article/technology-science-microsoft-corp-business-software-dd445694f34a6b7a0444db9988330229	4
Matthew Gault, <i>AI Trained on 4chan Becomes ‘Hate Speech Machine,’</i> Vice (June 7, 2022), https://www.vice.com/en/article/ai-trained-on-4chan-becomes-hate-speech-machine	22
<i>Microsoft and OpenAI Extend Partnership</i> , Official Microsoft Blog (Jan. 23, 2023), https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership	19
Microsoft Threat Intelligence, <i>AI Jailbreaks: What They Are and How They Can Be Mitigated</i> , Microsoft (June 4, 2024), https://www.microsoft.com/en-us/security/blog/2024/06/04/ai-jailbreaks-what-they-are-and-how-they-can-be-mitigated ..	19, 20
Mikaël Chelli et al., <i>Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis</i> , 26 J. Med. Internet Res. e53164 (2024).....	6
Mike Wendling, <i>The Saga of ‘Pizzagate’: The Fake Story that Shows How Conspiracy Theories Spread</i> , BBC News (Dec. 2, 2016), https://www.bbc.com/news/blogs-trending-38156985	22
Minhyeok Lee, <i>A Mathematical Investigation of Hallucination and Creativity in GPT Models</i> , 11 Mathematics 2320 (2023)	15
Nina Brown, <i>Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation</i> , 3 J. Free Speech L. 389 (2023)	14
OpenAI, <i>GPT-4 System Card</i> (Mar. 15, 2023), https://cdn.openai.com/papers/gpt-4-system-card.pdf	6, 7
OpenAI, <i>GPT-4 Technical Report</i> (Mar. 27, 2023), https://cdn.openai.com/papers/gpt-4.pdf	9, 19

Parth Sawhney, <i>How to Disable Bing Chat AI Responses in Bing Search</i> , All Things How (Apr. 7, 2023), https://allthings.how/how-to-disable-bing-chat-ai-responses-in-bing-search	4
Reece Rogers & Will Knight, <i>SearchGPT Is OpenAI's Direct Assault on Google</i> , Wired (July 25, 2024), https://www.wired.com/story/searchgpt-openai-search-engine-generative-ai	4
Reece Rogers, <i>How Google's AI Overviews Work, and How to Turn Them Off (You Can't)</i> , Wired (May 16, 2024), https://www.wired.com/story/google-ai-overviews-how-to-use-how-to-turn-off	4
S.M. Towhidul Islam Tonmoy et al., <i>A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models</i> , arXiv (Jan. 8, 2024), https://arxiv.org/abs/2401.01313v2	19
Sage Lazarro, <i>Big AI Thins Out the Competition as Startups Quit the Race to Build Large Language Models</i> , Fortune (Oct. 3, 2024), https://fortune.com/2024/10/03/openai-google-microsoft-amazon-llms-characterai	4
Shakked Noy & Whitney Zhang, <i>Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence</i> , 381 Science 187 (2023)	13
<i>Terms of Use</i> , OpenAI (Oct. 23, 2024), https://openai.com/policies/terms-of-use	5
Thomas Woodside & Helen Toner, <i>How Developers Steer Language Model Outputs: Large Language Models Explained, Part 2</i> , Ctr. for Sec. & Emerging Tech. (Mar. 8, 2024), https://cset.georgetown.edu/article/how-developers-steer-language-model-outputs-large-language-models-explained-part-2	16
Will Douglas Heaven, <i>Chatbots Could One Day Replace Search Engines. Here's Why that's a Terrible Idea.</i> , MIT Tech. Rev. (Mar. 29, 2022), https://www.technologyreview.com/2022/03/29/1048439/chatbots-replace-search-engine-terrible-idea	7
William H. Walters & Esther Isabelle Wilder, <i>Fabrication and Errors in the Bibliographic Citations Generated by ChatGPT</i> , 13 Sci. Reps. 14045 (2023)	6, 7
Yuntao Bai et al., <i>Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback</i> , arXiv (Apr. 12, 2022), https://arxiv.org/pdf/2204.05862	17
Zhengbao Jiang et al., <i>How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering</i> , 9 Transactions Ass'n for Computational Linguistics 962 (2021)	10
Ziwei Xu et al., <i>Hallucination Is Inevitable: An Innate Limitation of Large Language Models</i> , arXiv (Jan. 22, 2024), https://arxiv.org/abs/2401.11817	15

INTEREST OF *AMICUS CURIAE*

The Technology Law and Policy (“TLP”) Clinic at New York University School of Law is concerned with how technological advances drive legal, social, political, and economic change. The TLP Clinic has an interest in ensuring that settled law continues to serve the public good in the face of novel technologies. This case relates directly to that interest. The TLP Clinic submits this brief to assure that defamation law strikes the proper balance between permitting redress for reputational harms caused by generative artificial intelligence and protecting the speech interests in innovative communication and information technologies.

SUMMARY OF THE ARGUMENT

This case presents important and novel questions regarding how to apply defamation law to the content produced by generative artificial intelligence (“AI”) tools like OpenAI’s ChatGPT.¹ But the parties in this case offer all-or-nothing answers that would disrupt the proper balance between redressing reputational harm and creating breathing space for protected speech. In ruling on the pending motion for summary judgment, the Court should consider the core purposes of defamation law and conduct the fact-specific inquiry required to adapt the law to this new and increasingly widespread technology.

Defamation law should provide redress for the reputational harms caused by generative AI outputs. The world’s most powerful technology companies are increasingly embedding generative AI tools across our online lives, risking real and pervasive harms. OpenAI’s arguments to evade

¹ *Amicus* uses the term “generative AI” to refer both to generative AI tools that are trained to produce text and to those that are trained to produce realistic-looking images or other types of media. See Helen Toner, *What Are Generative AI, Large Language Models, and Foundation Models?*, Ctr. for Sec. & Emerging Tech. (May 12, 2023), <https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models>. When our discussion involves considerations or research specific to language-based generative AI tools, we use the narrower term “large language model” (“LLM”). *Id.*

liability in this case are thus misguided. Most significantly, despite the company’s suggestion to the contrary, reasonable users could believe that ChatGPT produces statements of fact in light of the full context surrounding a given statement, including the assurances of reliability provided by OpenAI in its marketing and by ChatGPT in its outputs. The Court should reject OpenAI’s attempt to obtain what would amount to categorical immunity from defamation claims for generative AI companies.

At the same time, the First Amendment should protect the interests that both generative AI companies and members of the public have in being able to provide and access generative AI tools. These interests have long been protected in defamation cases by the fault requirement, and the Court must now carefully and creatively apply this requirement to the unprecedented context of generative AI. In doing so, the Court should look to established principles of products liability and agency law to conduct a fact-specific inquiry into whether the company that developed and deployed the generative AI tool has met the requisite level of fault.

No court has yet directly grappled with the difficult challenges posed for defamation law when a corporation develops and deploys a generative AI tool capable of making false, apparently factual claims about specific individuals. The Court’s analysis will help shape the development of this area of the law—both in Georgia and throughout the country.

ARGUMENT

I. Defamation law should protect against reputational harms caused by generative AI outputs.

A handful of powerful technology companies currently dominate the market for generative AI tools. These companies are installing generative AI tools across a wide range of online services through unilateral contracts of adhesion that members of the general public have no meaningful

opportunity to negotiate. As generative AI tools become increasingly commonplace, they carry significant risk of producing outputs that, despite their factual inaccuracy, users may rely upon.

The Court should therefore reject OpenAI's argument that no reasonable user of ChatGPT could ever believe its outputs constitute accurate statements of fact because the company issues certain disclaimers regarding the reliability of its generative AI tool. In defamation cases, the determination of whether a statement is an actionable statement of fact turns on all the relevant circumstances, and disclaiming language is only one part of the analysis. In making this determination here, the Court should take into account the ways OpenAI has generally represented and marketed ChatGPT as a reliable tool as well as ChatGPT's specific pattern of contradicting its own statements regarding its limitations in this case.

A. Powerful technology companies are increasingly embedding generative AI tools throughout the internet, creating risks of real and widespread harms.

The market for generative AI tools is highly concentrated, and ChatGPT is at the forefront of this concentration of power. According to one report, OpenAI's tools represent 76% of the market share for large language models, the types of generative AI that are trained on vast amounts of linguistic data to recognize and generate human language.² OpenAI estimates that ChatGPT has over 200 million weekly active users.³ Training a competitive LLM is an extremely costly process, requiring on the order of tens of millions of dollars in investment.⁴ Because of these costs, almost

² See Anton Korinek & Jai Vipra, *Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT* 7 (Ctr. on Regul. & Mkts. at Brookings, Working Paper No. 9, 2023), <https://arxiv.org/pdf/2311.01550>.

³ See Benj Edwards, *ChatGPT Hits 200 Million Active Weekly Users, but How Many Will Admit Using It?*, *Ars Technica* (Aug. 30, 2024), <https://arstechnica.com/information-technology/2024/08/chatgpt-hits-200-million-active-weekly-users-but-how-many-will-admit-using-it>.

⁴ See Katharina Buchholz, *The Extreme Cost of Training AI Models*, *Forbes* (Aug. 26, 2024), <https://www.forbes.com/sites/katharinabuchholz/2024/08/23/the-extreme-cost-of-training-ai-models>.

all generative AI tools are based on “foundation models” produced by a small number of Big Tech companies.⁵ And as the costs to compete only continue to rise, smaller players are exiting the market, leaving the development of generative AI tools almost entirely in the hands of a very small number of companies.⁶

These technology companies are increasingly embedding generative AI tools into every aspect of the online environment. Even for people who choose not to sign up to use LLMs, generative AI is increasingly a part of their online life, with Big Tech companies integrating LLMs into most major search engines: Both Google and Bing search engines already frequently display AI-generated content at the top of their search results.⁷ Although Bing allows users to opt out of this AI-driven search engine functionality, Google does not.⁸ These uses of LLMs make generative AI a routine part of online life, which in turn encourages internet users to view LLM outputs as reliable sources of gathering information. OpenAI itself recently released its own search engine, SearchGPT, which will accelerate the sheer amount of generative AI outputs while promoting the notion that LLM programs provide reliable facts.⁹ The use of LLMs as a tool for fact-finding and research is not far-fetched speculation, but a reality that has crept into the information ecosystem

⁵ See Amba Kak & Sarah Myers West, *ChatGPT and More: Large Scale AI Models Entrench Big Tech Power*, AI Now Inst. (Apr. 11, 2023), <https://ainowinstitute.org/publication/large-scale-ai-models>.

⁶ See Sage Lazarro, *Big AI Thins Out the Competition as Startups Quit the Race to Build Large Language Models*, Fortune (Oct. 3, 2024), <https://fortune.com/2024/10/03/openai-google-microsoft-amazon-llms-characterai>.

⁷ See Matt O’Brien, *Microsoft Bakes ChatGPT-like Tech into Search Engine Bing*, Associated Press (Feb. 7, 2023), <https://apnews.com/article/technology-science-microsoft-corp-business-software-dd445694f34a6b7a0444db9988330229>.

⁸ See Parth Sawhney, *How to Disable Bing Chat AI Responses in Bing Search*, All Things How (Apr. 7, 2023), <https://allthings.how/how-to-disable-bing-chat-ai-responses-in-bing-search>; Reece Rogers, *How Google’s AI Overviews Work, and How to Turn Them Off (You Can’t)*, Wired (May 16, 2024), <https://www.wired.com/story/google-ai-overviews-how-to-use-how-to-turn-off>.

⁹ See Reece Rogers & Will Knight, *SearchGPT Is OpenAI’s Direct Assault on Google*, Wired (July 25, 2024), <https://www.wired.com/story/searchgpt-openai-search-engine-generative-ai>.

for most internet users who depend on mainstream search engines, even those who have never accessed or registered to use generative AI tools like ChatGPT.

Companies like OpenAI are also offering or plainly imposing these new tools on unilateral terms that users have no meaningful opportunity to negotiate or reject. These terms of service resemble, at best, standardized contracts of adhesion “offered on a ‘take it or leave it’ basis and under such conditions that a consumer cannot obtain the desired product or service except by acquiescing in the form contract.” *Hosp. Auth. of Houston Cnty. v. Bohannon*, 272 Ga. App. 96, 98–99 (2005). But these companies should not be able to exploit their power in the marketplace, and the power of these new tools, to force users to contract away important rights and responsibilities.¹⁰ OpenAI’s Terms of Use for ChatGPT, for example, attempt to hold users solely responsible for the outputs of its generative AI tool.¹¹ Powerful actors attempting to unilaterally impose form contracts is no new phenomenon, but with a new communication technology like generative AI, relying on these terms of adhesion to shift all responsibility over outputs onto consumers would harm not only users but also third parties and the public interest.¹² Because generative AI users cannot meaningfully negotiate their contract terms, and third parties who stand to be reputationally harmed by erroneous generative AI outputs are not even at the table to consent

¹⁰ David A. Hoffman, *Defeating the Empire of Forms*, 109 Va. L. Rev. 1367, 1375 (2023) (“Forms are full of clauses that exclude tort remedies, [and] waive property standards When coupled with procedural devices that make it harder to vindicate such small-stakes individual harms in court, small-stakes forms off-load risk to the public.”).

¹¹ *See Terms of Use*, OpenAI (Oct. 23, 2024), <https://openai.com/policies/terms-of-use> (“You are responsible for Content [(i.e., ChatGPT inputs and outputs)], including ensuring that it does not violate any applicable law or these Terms.”).

¹² *See* Aditi Bagchi, *Other People’s Contracts*, 32 Yale J. on Regul. 211, 243 (2015) (“Third party interests are one element in the interpretive canon that properly informs a choice among reasonable meanings of an ambiguous term.”).

or negotiate, terms of service by generative AI companies should be viewed with suspicion by courts when raised as a shield against defamation liability.

Despite their pervasive presence online, LLMs are not reliable sources of truth. Research has shown that generative AI programs like ChatGPT produce massive amounts of untrue outputs by “hallucinating” information that is both factually unfounded and made up (that is, not present in the model’s training data).¹³ When asked to provide evidence to support its outputs, ChatGPT often resorts to outputting even more unsubstantiated hallucinations.¹⁴ OpenAI’s own research shows that the GPT system continues to hallucinate across older and newer versions of the model.¹⁵ Hallucination is not merely a bug, but part of the fundamental functioning of LLMs, which are

¹³ See William H. Walters & Esther Isabelle Wilder, *Fabrication and Errors in the Bibliographic Citations Generated by ChatGPT*, 13 *Sci. Reps.* 14045, 4 (2023) (finding fabricated citations in 55% of GPT-3.5 outputs and 18% of GPT-4 outputs); Mikaël Chelli et al., *Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis*, 26 *J. Med. Internet Res.* e53164, 5 (2024) (finding hallucinated medical literature in 39.6% of outputs by GPT-3.5, 28.6% by GPT-4, and 91.4% by Bard); Junyi Li et al., *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models*, Proc. 2023 Conf. on Empirical Methods Nat. Language Processing 6449, 6453 (evaluating 35,000 output samples and finding that ChatGPT hallucinations appeared in 19.5% of its outputs).

¹⁴ See Guido Zuccon et al., *ChatGPT Hallucinates when Attributing Answers*, Proc. Ann. Int’l ACM SIGIR Conf. on Rsch. & Dev. Info. Retrieval Asia Pac. Region 46, 49 (2023) (finding that 86% of references outputted by ChatGPT did not exist).

¹⁵ See OpenAI, *GPT-4 System Card* 41, 46, 59–60 (Mar. 15, 2023), <https://cdn.openai.com/papers/gpt-4-system-card.pdf>.

designed to mimic texts, not to process their substantive content.¹⁶ The technology companies that offer generative AI tools recognize that hallucinations in LLM outputs will not be easily fixed.¹⁷

Hallucinations, combined with increased user trust in generative AI tools, can be “particularly harmful” as users become reliant on them.¹⁸ As OpenAI itself has acknowledged, “[h]allucinations can become more dangerous as models become more truthful, as users build trust in the model when it provides truthful information in areas where they have some familiarity.”¹⁹ Studies suggest that users tend to trust in a technology the more they depend on that technology across different situations and uses.²⁰ In the current barrage of roll-outs of generative AI tools, consumers are repeatedly exposed to generative AI-driven tools built in to search engines and consumer service chatbots and are being told that they are useful, reliable, and even empathetic. This builds user trust in generative AI tools. As a result, even when users may be generally aware that some information provided by generative AI tools could be untrue, psychological phenomena like the mere-exposure effect or trust in generative AI tools are likely to result in people intuitively believing their outputs.²¹

¹⁶ See Walters & Wilder, *supra* note 13, at 4; Will Douglas Heaven, *Chatbots Could One Day Replace Search Engines. Here’s Why that’s a Terrible Idea.*, MIT Tech. Rev. (Mar. 29, 2022), <https://www.technologyreview.com/2022/03/29/1048439/chatbots-replace-search-engine-terrible-idea>; Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, Proc. 2021 ACM Conf. on Fairness, Accountability & Transparency 610, 610 (“[Language models] are not performing natural language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form.”).

¹⁷ See Matt O’Brien, *Chatbots Sometimes Make Things Up. Is AI’s Hallucination Problem Fixable?*, Associated Press (Aug. 1, 2023), <https://apnews.com/article/artificial-intelligence-hallucination-chatbots-chatgpt-falsehoods-ac4672c5b06e6f91050aa46ee731bcf4>.

¹⁸ OpenAI, *supra* note 15, at 46.

¹⁹ *Id.*

²⁰ See Francesca Cabiddu et al., *Why Do Users Trust Algorithms? A Review and Conceptualization of Initial Trust and Trust over Time*, 40 Eur. Mgmt. J. 685, 689 (2022).

²¹ See Celeste Kidd & Abeba Birhane, *How AI Can Distort Human Beliefs*, 380 Science 1222, 1222 (2023).

B. Reasonable users could believe generative AI outputs constitute statements of fact.

The Court should reject OpenAI’s argument that no reasonable user of ChatGPT could believe that its outputs constitute statements of fact. *See* Def.’s Mem. Supp. Mot. Summ. J. 14–15. To state a claim for defamation, a plaintiff must point to a specific false statement of fact made by the defendant. Whether the statement at issue could “‘reasonably [be] interpreted as stating actual facts’ about an individual” is a fact-dependent inquiry that turns on all the relevant circumstances, including the language of the statement and its context. *Milkovich v. Lorain Journal Co.*, 497 U.S. 1, 9, 20 (1990) (quoting *Hustler Mag., Inc. v. Falwell*, 485 U.S. 46, 50 (1988)). In conducting this inquiry, Georgia courts have looked to the “context of the entire writing in which the [statement] appears,” *Gast v. Brittain*, 277 Ga. 340, 341 (2003), and “examine[d] the statement in its totality in the context in which it was uttered or published,” *Bollea v. World Championship Wrestling, Inc.*, 271 Ga. App. 555, 558 (2005).

While acknowledging the context-specific nature of this inquiry, OpenAI’s motion for summary judgment focuses on only *one* contextual factor that the Court should use to evaluate the meaning of ChatGPT’s outputs: disclaimers placed underneath ChatGPT’s input box and embedded in its Terms of Use. *See* Def.’s Mem. Supp. Mot. Summ. J. 13–15. Based solely on this factor, OpenAI argues that “no reasonable person” could conclude that a statement of a factual nature from ChatGPT means what it apparently asserts. *Id.* at 15.

But these disclaimers should merely be one part of the analysis and are not dispositive in and of themselves. Georgia courts have considered similar disclaimers and found them insufficient in light of the plain meaning of the allegedly defamatory statements at issue as well as the broader context in which the statements were made. *See, e.g., Harcrow v. Struhar*, 236 Ga. App. 403, 404 (1999) (holding that defendant’s disclaimer that “I’m not saying that [plaintiffs] are responsible

for this atrocious act” did not overcome the plain meaning of the accusation). And the U.S. Supreme Court has held that merely couching a statement with qualifiers such as “in my opinion” does not generally dispel the factual implications that make an otherwise defamatory statement actionable. *Milkovich*, 497 U.S. at 19 (“[It] would be destructive of the law of libel if a writer could escape liability for accusations of crime simply by using, explicitly or implicitly, the words ‘I think.’”) (quoting *Cianci v. New Times Publ’g Co.*, 639 F.2d 54, 64 (2d Cir. 1980)).²²

Instead, the Court should consider the full linguistic and social context of the outputs at issue in determining whether the user—here, Mr. Riehl—could have reasonably understood the outputs as stating facts. In this case, this context includes the ways in which OpenAI has represented and marketed ChatGPT to the public and users. OpenAI does not market ChatGPT or its other generative AI tools as fiction machines. For example, it has promoted its recent launch of GPT-4 for its “greater accuracy, thanks to its broader general knowledge and problem-solving abilities.”²³ Moreover, greater reliance is precisely what generative AI companies encourage from their users as companies like OpenAI emphasize their tools’ high performance on things like various standardized tests.²⁴ As discussed in Part I.A, *supra*, OpenAI has even released “SearchGPT,” a service built on top of ChatGPT that acts as a search engine. It would thwart the purposes of defamation law to allow a powerful technology company to market its tools as accurate

²² See also Eugene Volokh, *Large Libel Models? Liability for AI Outputs*, 3 J. Free Speech L. 489, 500 (2023) (“No newspaper can immunize itself from libel lawsuits for a statement that ‘Our research reveals that John Smith is a child molester’ . . . by putting a line on the front page, ‘Warning: We may sometimes publish inaccurate information[.]’”).

²³ *GPT-4 Is OpenAI’s Most Advanced System, Producing Safer and More Useful Responses*, OpenAI (last accessed Dec. 3, 2024), <https://openai.com/index/gpt-4>.

²⁴ OpenAI, *GPT-4 Technical Report* 5–6 (Mar. 27, 2023), <https://cdn.openai.com/papers/gpt-4.pdf>.

and reliable while simultaneously claiming that no reasonable person could ever believe the outputs of those tools.

Equally relevant context is ChatGPT’s own outputs surrounding the allegedly defamatory statements, including erratic behavior and overconfidence, both of which are well-established features of generative AI tools.²⁵ This is underscored by the transcript in the present case. ChatGPT initially stated that it would be unable to complete Mr. Riehl’s requests. *See* Def.’s Answer to Pl.’s Am. Compl. 91 (Ex. 8) (“I’m sorry, but as an AI language model, I do not have access to the internet and cannot read or retrieve any documents.”).²⁶ If the conversation had ended there, a reasonable user *may* have inferred that every subsequent output was a flight of fancy containing no factual assertions whatsoever.²⁷ However, it quickly changed tack and informed Mr. Riehl that “Yes, I can read the document you provided,” later referring to its own hallucinated document as a “genuine legal complaint.” *Id.* at 92, 101. This sort of behavior is common with LLMs. These models are inept at estimating their own level of confidence in their outputs and frequently contradict themselves, as they operate on models of linguistic probabilities rather than reasoning.²⁸ This can lead to a model seeming to acknowledge its own limitations in one output, only to confidently assert that it had overcome those limitations in the next. Given these behaviors, combined with the newness of the technology and the lack of settled expectations regarding how

²⁵ *See, e.g.,* Zhengbao Jiang et al., *How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering*, 9 Transactions Ass’n for Computational Linguistics 962, 968 (2021) (“We found that [language models] tend to be over-confident about cases they do not know, as shown . . . [by the fact] that most predictions have aggressive confidence being close to 0 or 1.”).

²⁶ OpenAI has filed the relevant portions of the transcript as a sealed exhibit to its motion for summary judgment. *See* Def.’s Mot. Summ. J. (Ex. C) (filed under seal).

²⁷ *See* Volokh, *supra* note 22, at 500–01 (“To be sure, if a disclaimer actually describes something as *fiction*, or as parody or a hypothetical (both forms of fiction), that may well preclude defamation liability.”).

²⁸ *See* Jiang et al., *supra* note 25.

LLMs behave, it beggars belief to insist that “no reasonable user” could be misled into treating a literal, non-hyperbolic description as a statement of fact. Nevertheless, OpenAI would have the Court find that ChatGPT outputs could never be considered statements of fact, because the company issues disclaimers regarding the factual reliability of its generative AI tool. A categorical rule that accepts powerful AI companies’ unilateral disclaimers as dispositive would create unprecedented blanket immunity for all defamatory generative AI outputs. Such a rule would run contrary to settled law that requires courts to grapple with the full context surrounding generative AI outputs, including the generative AI companies’ own assurances of accuracy and the contradictory statements made by generative AI tools in their outputs.

II. Defamation plaintiffs should be required to allege specific unreasonably risky conduct taken by generative AI companies.

AI companies have an interest in offering generative AI tools to the public, and the public has an interest in being able to access them. In defamation cases, the fault requirement serves to protect the First Amendment interests of both speakers and listeners. But this fault requirement has always assumed that speech is the product of conscious human choice, not the indirect result of decision-making by a developer or deployer of a communicative tool. To adapt the fault requirement to the context of generative AI outputs—and to safeguard the First Amendment interests that both generative AI companies and users have in the creation and use of generative AI tools—courts should therefore creatively borrow from other bodies of law that have dealt with analogous factual circumstances. In particular, courts can look to principles drawn from products liability and agency law to determine whether a company has demonstrated the requisite level of fault with respect to an allegedly defamatory generative AI output. Products liability frameworks can be particularly useful for evaluating a developer’s pre-deployment design decisions, while agency law frameworks can help evaluate the post-deployment monitoring of a generative AI tool.

A. Imposing a fault requirement protects both generative AI companies’ and the public’s First Amendment interests in generative AI tools.

The First Amendment imposes a fundamental requirement on all defamation claims: a plaintiff must be able to demonstrate some form of fault on the part of a defendant. *See Gertz v. Robert Welch, Inc.*, 418 U.S. 323, 347 (1974) (“We hold that, *so long as they do not impose liability without fault*, the States may define for themselves the appropriate standard of liability for a publisher or broadcaster of defamatory falsehood injurious to a private individual.” (emphasis added)). This fault requirement has been described by the Supreme Court as providing the necessary “breathing space” that allows speakers to exercise their First Amendment rights without facing a chilling threat of litigation. *See id.* at 342. And generative AI companies likely do have some form of protected First Amendment interest in being able to develop, deploy, and market their tools.²⁹

But even as the status of generative AI companies as First Amendment speakers remains unsettled, the First Amendment also protects the public’s “right to receive information and ideas.” *Stanley v. Georgia*, 394 U.S. 557, 564 (1969). This “right to receive information” is rooted in the fact that members of the public cannot meaningfully exercise their own free speech without being

²⁹ Courts have not yet determined the exact nature of the First Amendment interest that generative AI developers or deployers have in their tools. ChatGPT outputs may be a form of “corporate speech.” *See Citizens United v. Fed. Election Comm’n*, 558 U.S. 310, 319 (2010) (“The Government may regulate corporate political speech through disclaimer and disclosure requirements, but it may not suppress that speech altogether.”). OpenAI’s attempts to curate and monitor the types of outputs that ChatGPT produces may amount to protected “editorial” decisions. *See Moody v. NetChoice, LLC*, 144 S. Ct. 2383, 2403 (2024) (“The government may not, in supposed pursuit of better expressive balance, alter a private speaker’s own editorial choices about the mix of speech it wants to convey.”). Or the ability to offer some form of access to ChatGPT may be a form of speech on the part of the programmers who developed the model, just as courts have held that the First Amendment protects the open publication of computer code. *See, e.g., Universal City Studios, Inc. v. Corley*, 273 F.3d 429, 448 (2d Cir. 2001) (“Limiting First Amendment protection of programmers to descriptions of computer code (but not the code itself) would impede discourse among computer scholars . . .”).

assured of access to a vigorous exchange of ideas. *See Bd. of Educ. v. Pico*, 457 U.S. 853, 867 (1982). Despite generative AI’s risks, the public has a significant interest in being able to continue accessing these tools and their numerous potentially valuable uses that contribute to the exchange of ideas, which range from generating computer code to even acting as a mediator that can improve group reasoning.³⁰ Experimental results suggest that the use of ChatGPT can improve the productivity of writers tasked to draft press releases, reports, or “delicate emails.”³¹

Thus, imposing the traditional fault requirement of defamation law is not only speaker-protective, but also protective of the public’s right to receive information insofar as it allows for communication “without previous restraint or fear of subsequent punishment.” *First Nat’l Bank of Boston v. Bellotti*, 435 U.S. 765, 775–76 (1978) (rejecting the question of whether corporations “have” First Amendment rights to focus on the question of whether the speech at issue is of the type the First Amendment was meant to protect). The fault requirement should therefore continue to play a critical function in protecting these First Amendment interests in the context of generative AI tools.

B. Evaluating the fault of generative AI companies requires creatively borrowing from other bodies of law.

Given the First Amendment interests that both generative AI companies and the public have in AI-generated speech, the Court should carefully adapt the fault element of a defamation claim to account for the unprecedented factual context of generative AI tools. Although the relevant standards of fault—negligence and actual malice—are well established, defamation case law can only be so helpful in applying these standards to the context of generative AI tools, which

³⁰ *See* Helena Kudiabor, *AI Tool Helps People with Opposing Views Find Common Ground*, *Nature* (Oct. 17, 2024), <https://www.nature.com/articles/d41586-024-03424-z>.

³¹ *See* Shakked Noy & Whitney Zhang, *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence*, 381 *Science* 187, 187 (2023).

in important respects do not resemble the traditional human “speakers” and “publishers” that the First Amendment has long protected.

Other areas of law, however, have considered analogous situations. Products liability law, for example, extensively addresses when harms caused by a product’s design or warnings can be attributed to a designer’s negligence. *See* Restatement (Third) of Torts: Prods. Liab. § 2 cmt. a (Am. L. Inst. 1998) (stating that design defect standards are meant to “achieve the same general objectives as . . . liability predicated on negligence”). And agency law exists for the purpose of determining when one actor’s conduct may be imputed to another due to the existence of a principal-agent relationship. *See* Restatement (Third) of Agency intro. (Am. L. Inst. 2006).

Courts adapting defamation law to address generative AI speech can look to concepts found in products liability and agency law to evaluate whether generative AI companies are at fault at two stages where they could otherwise intervene to reduce the risk of defamatory outputs. In the first, pre-deployment stage, companies design, train, and refine their model. In doing so, they may fail to adopt reasonable alternative design decisions that would have mitigated the harm of foreseeable defamatory outputs—the type of inquiry for which design defect cases provide guidance.³² In the second, post-deployment stage, companies monitor how users interact with the tools, making continual adjustments to the service and revoking access from users who appear to be abusing it.³³ As at the first stage, companies may fail to take reasonable precautions that would

³² *See, e.g.*, Nina Brown, *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*, 3 J. Free Speech L. 389, 410–14 (2023) (“In the context of speech harms caused by chatbots, a design defect could exist if the model was designed in a way that made it likely to generate defamatory statements.”).

³³ *See Disrupting Malicious Uses of AI by State-Affiliated Threat Actors*, OpenAI (Feb. 14, 2024), <https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors> (demonstrating such a capability).

mitigate the harms of potential defamatory outputs, though at this stage the failure to do so more closely resembles an employer’s negligent supervision of their employee.

In the instant case, the Court should consider whether Mr. Walters has identified specific instances of unreasonably risky conduct, or whether he has merely shown that OpenAI had a general awareness of hallucination risks before deploying ChatGPT. *See* Third Am. Compl. ¶¶ 35–38, 44–45 (referring to press releases and articles about hallucination risks). Many computer scientists believe that hallucinations are an inherent feature of generative AI systems that can be mitigated but not fully eliminated.³⁴ If courts were to find that it is negligent to deploy a generative AI tool merely because it carries some risk of hallucination, and it is impossible to fully eliminate the risk of hallucination from generative AI models, then merely deploying any generative AI model would always be considered negligent. Instead of accepting this theory, the Court should consider whether the defamation claim presented identifies an alternative reasonable design or specific error in supervision that would justify a finding of negligence consistent with principles of products liability or agency law. And if the Court decides that Mr. Walters is a public figure, it should conduct a fact-specific actual malice inquiry into OpenAI’s conduct in developing and monitoring ChatGPT without assuming, as OpenAI urges, that actual malice can never be imputed to a “computer output.” Def.’s Mem. Supp. Mot. Summ. J. 22 n.3.

1. Products liability law can be used to evaluate the negligence of generative AI companies with regard to pre-deployment design decisions.

Products are considered defective in design when “the foreseeable risks of harm posed by the product could have been reduced or avoided by the adoption of a reasonable alternative

³⁴ *See* Minhyeok Lee, *A Mathematical Investigation of Hallucination and Creativity in GPT Models*, 11 *Mathematics* 2320, 2335 (2023) (concluding that “hallucinations may be an intrinsic characteristic of GPT models”); *accord* Ziwei Xu et al., *Hallucination Is Inevitable: An Innate Limitation of Large Language Models*, arXiv (Jan. 22, 2024), <https://arxiv.org/abs/2401.11817>.

design . . . and the omission of the alternative design renders the product not reasonably safe.” Restatement (Third) of Torts: Prods. Liab. § 2 (Am. L. Inst. 1998); *see also Jones v. NordicTrack, Inc.*, 274 Ga. 115, 118 (2001) (describing the reasonable alternative design standard as discussed in the Restatement (Third) as the “heart” of a design defect case). To make out a claim that a product was defectively designed, plaintiffs must do more than merely point out a danger inherent in a product, which is only one factor in the risk-utility analysis. *See Banks v. ICI Americas, Inc.*, 264 Ga. 732, 734 (1994). They must identify a reasonable alternative design, one which “would have made the product safer than the original design and was a marketable reality and technologically feasible” at the time the product was designed. *Id.* at 736. An important consideration in determining whether a reasonable alternative design was “practicable” at the time of manufacture is often the industry custom and the designs used by competitors. *See* Restatement (Third) of Torts: Prods. Liab. § 2 cmt. d (Am. L. Inst. 1998).

AI developers make an enormous number of choices in designing a generative AI tool, from curating data, to choosing how much computational power to expend on model training, to deciding whether or how to “fine-tune” models in order to make them more likely to behave in certain ways. Many of these design decisions impact a model’s likelihood to generate hallucinations.³⁵ Mitigation techniques that can reduce the risk of hallucinations are constantly being innovated. For instance, in 2022, generative AI company Anthropic began using a method dubbed “reinforcement learning from human feedback” to improve model outputs by showing a

³⁵ *See* 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://cset.georgetown.edu/article/how-developers-steer-language-model-outputs-large-language-models-explained-part-2> (discussing methods by which generative AI developers “fine-tune” models to be less likely to engage in certain types of undesired behavior).

model many human-annotated examples of “helpful” outputs.³⁶ And in 2024, OpenAI announced that its new “o1” model improved on the state of the art in terms of accuracy and reliability by using “chain of thought” reasoning, wherein the model essentially speaks to itself to spot flaws in its own reasoning before providing an output to the user.³⁷ Where plaintiffs can demonstrate that a developer failed to take certain mitigating measures in designing and training a generative AI model—especially where those measures have become an industry norm—they may plausibly have identified a reasonable alternative design, thus showing the developer’s negligence.

This analogy to products law is not without some motivation in traditional defamation standards. Refusal to follow industry custom in integrating mitigating design elements into a generative AI tool’s design is, after all, analogous to a publisher’s failure to follow professional custom in investigating claims prior to publication—a consideration that the Restatement (Second) of Torts highlights when evaluating the negligence of publishers in defamation cases. *See* Restatement (Second) of Torts § 580B cmt. g (Am. L. Inst. 1977); *see also* *Curtis Publ’g Co. v. Butts*, 388 U.S. 130, 158 (1967) (finding publisher liability given “an extreme departure from the standards of investigation and reporting ordinarily adhered to by responsible publishers”). This approach to demonstrating negligence, in other words, has roots in defamation law, but borrows from products liability law in order to appropriately shift the focus from “standards of investigation” to standards of designing and developing a generative AI tool.³⁸

³⁶ *See* Yuntao Bai et al., *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*, arXiv (Apr. 12, 2022), <https://arxiv.org/pdf/2204.05862>.

³⁷ *See* *Learning to Reason with LLMs*, OpenAI (Sept. 12, 2024), <https://openai.com/index/learning-to-reason-with-llms>.

³⁸ Requiring plaintiffs to allege a reasonable alternative design may increase the cost of some defamation suits against generative AI companies, as plaintiffs may need to identify experts to testify regarding industry custom in designing generative AI tools. But as OpenAI points out, this is similar to many traditional defamation cases, where “[c]ustoms and practices within the profession are relevant in applying the negligence standard, which is, to a substantial degree, set

In the immediate case, the Court should question whether liability could be established based only on evidence that OpenAI was aware of the general risk of hallucinations, as such a low bar for finding defamation fault would raise serious First Amendment concerns by heavily curtailing generative AI companies' ability to offer their tools to the public. In particular, the Court should query whether Mr. Walters has identified a design decision which, if implemented, would have prevented ChatGPT from making the outputs about him at issue. Requiring a plaintiff to identify a reasonable alternative design—one which was a “marketable reality and technologically feasible” when the model was first designed, *Banks*, 264 Ga. at 736—would be one way for a court to find liability without seriously jeopardizing the First Amendment's fault requirement.

2. Agency law can be used to evaluate the negligence of generative AI companies with regard to post-deployment oversight decisions.

Principals are considered negligent in their supervision of agents when they “conduct[] an activity through an agent,” the agent causes harm to a third party, and “the harm was caused by the principal's negligence in selecting, training, retaining, supervising, or otherwise controlling the agent.” Restatement (Third) of Agency § 7.05 (Am. L. Inst. 2006). In addition, a claim of negligent supervision requires demonstrating that a principal had the legal right to control the “the time, manner, and method of [the agent's] day-to-day operations.” *DaimlerChrysler Motors Co. v. Clemente*, 294 Ga. App. 38, 44 (2008).

AI companies can take a number of post-deployment actions to mitigate the risks of their tools in ways that resemble employee supervision. Microsoft—which has a close relationship with

by the profession itself,’ and such proof ‘would normally come from an expert who has been shown to be qualified on the subject.’” Def.’s Mem. Supp. Mot. Summ. J. 24 (citing *Gettner v. Fitzgerald*, 297 Ga. App. 258, 265 n.8 (2009)).

OpenAI and was instrumental in developing ChatGPT³⁹—has said that “AI language models can be likened to employees who are enthusiastic and knowledgeable but lack the judgment, context understanding, and adherence to boundaries that come with experience and maturity in a business setting.”⁴⁰ The actions that generative AI companies can take to “supervise” their employee-chatbots can include “prompt engineering” tactics that alter user queries in ways that reduce the likelihood of harmful outputs.⁴¹ Or companies may develop and update filters that screen out content that is likely to be harmful before returning it to a user.⁴² This filtering may include the ability to filter out content regarding specific individuals, or specific claims about them—a measure that OpenAI appears to have already taken in a number of instances.⁴³ And companies may monitor *user* behavior and revoke the access of users who appear to be clearly abusing the service, for instance by using it for the purpose of generating disinformation.⁴⁴ A failure to implement these “supervisory” mitigations post-deployment could justify a finding of negligence

³⁹ See *Microsoft and OpenAI Extend Partnership*, Official Microsoft Blog (Jan. 23, 2023), <https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership>.

⁴⁰ Microsoft Threat Intelligence, *AI Jailbreaks: What They Are and How They Can Be Mitigated*, Microsoft (June 4, 2024), <https://www.microsoft.com/en-us/security/blog/2024/06/04/ai-jailbreaks-what-they-are-and-how-they-can-be-mitigated>.

⁴¹ See S.M. Towhidul Islam Tonmoy et al., *A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models*, arXiv 2–7 (Jan. 8, 2024), <https://arxiv.org/abs/2401.01313v2>.

⁴² See OpenAI, *supra* note 24, at 66.

⁴³ See, e.g., Ashley Belanger, *Will ChatGPT’s Hallucinations Be Allowed to Ruin Your Life?*, Ars Technica (Oct. 23, 2023), <https://arstechnica.com/tech-policy/2023/10/will-chatgpts-hallucinations-be-allowed-to-ruin-your-life> (reporting that, after Australian mayor Brian Hood accused ChatGPT of producing defamation about him, “the company had filtered the false statements about Hood from ChatGPT”); Benj Edwards, *Certain Names Make ChatGPT Grind to a Halt, and We Know Why*, Ars Technica (Dec. 2, 2024), <https://arstechnica.com/information-technology/2024/12/certain-names-make-chatgpt-grind-to-a-halt-and-we-know-why> (discussing multiple individuals on which OpenAI appears to have placed filters).

⁴⁴ See *Disrupting Deceptive Uses of AI by Covert Influence Operations*, OpenAI (May 30, 2024), <https://openai.com/index/disrupting-deceptive-uses-of-AI-by-covert-influence-operations> (demonstrating monitoring and access revocation capabilities).

on the part of an generative AI company. The key issue is whether a company had the capability to control the “time, manner, and method” of its model’s outputs, *DaimlerChrysler*, 294 Ga. App. at 44, and whether it failed to take steps to do so.

Here too, the analogy to negligent supervision has some grounding in standard defamation law, which occasionally grapples with suits imputing the speech of employees to their employers. Georgia courts have required that in these cases, a plaintiff must show that an employer had “some control” over a particular defamatory statement made by an employee. *See Mullinax v. Miller*, 242 Ga. App. 811, 814 (2000). For some outputs, whether or not the generative AI company was able to exert control over the output may be a disputed question of fact, as when users deliberately circumvent filters and controls intended to limit a model’s likelihood of making defamatory outputs.⁴⁵ Whether or not a claim that sounds in negligent supervision should succeed in these cases should depend on factual questions regarding the capacity of the generative AI company to control its model’s outputs in the face of such attacks.

The Court should consider whether the record here includes any evidence that OpenAI failed to take available steps to better supervise ChatGPT’s output and whether OpenAI could have controlled the outputs at issue in this case or prevented ChatGPT from making them. Given the tendency of LLMs to hallucinate rather than disappoint their user, and the imperfect state of the art with respect to controlling this problem, it is unclear whether OpenAI could have controlled the outputs at issue in this case or prevented ChatGPT from making them.

⁴⁵ *See generally* Microsoft Threat Intelligence, *supra* note 40 (discussing “jailbreaking” of LLMs and some mitigation tactics, as well as their limitations).

3. Both products liability and agency law can be used to evaluate particularly egregious conduct that may raise an inference of actual malice.

Generative AI companies may engage in conduct that also meets the heightened actual malice standard. Actual malice has traditionally been demonstrated in rare cases involving public figures or officials whenever a defendant is shown to have made a statement “with reckless disregard of whether it was false or not.” *N.Y. Times Co. v. Sullivan*, 376 U.S. 254, 280 (1964). “Reckless disregard,” the Supreme Court has acknowledged, does not have “one infallible definition,” and evaluating it requires a very fact-specific evaluation of “concrete cases.” *St. Amant v. Thompson*, 390 U.S. 727, 730–31 (1968). Evaluating actual malice, in other words, necessarily involves a fresh evaluation of the circumstances of each case.

OpenAI’s motion for summary judgment, however, points to dicta in past cases that assumed a human speaker behind every statement and elevates these dicta into a formal requirement for a claim of actual malice. *See* Def.’s Mem. Supp. Mot. Summ. J. 23 (“When a plaintiff sues an organization like OpenAI for defamation, the plaintiff must identify *specific* individuals within the organization who acted with actual malice.”). In *Sullivan*, the Supreme Court required that a showing of actual malice must be “brought home” to the persons “having responsibility” for the publication. 376 U.S. at 287. But in making this statement, the Court was making the limited point that prior news stories published by the *Times* could not be used to suggest corporate malice where the employees who approved the advertisement at issue had no knowledge of those prior stories. *Id.* From this remark and a scattering of federal appellate cases, OpenAI has manufactured a requirement that Mr. Walters identify a specific employee who “drafted the output, reviewed it, or . . . knew about it” at the time ChatGPT produced the statement. Def.’s Mem. Supp. Mot. Summ. J. 23. Accepting this argument would lead inexorably to the conclusion that OpenAI briefly teases in a footnote: that a generative AI company could *never* be liable for the statements

of its tools regarding public figures or officials simply because those statements are a “computer output.” *See* Def.’s Mem. Supp. Mot. Summ. J. 22 n.3. Such a conclusion would result in the same dismal outcome as OpenAI’s arguments regarding its Terms of Use: a complete frustration of defamation law simply because speech is produced by a generative AI tool.

Yet there are scenarios where a developer’s conduct could plausibly suggest reckless disregard for the truth. And although there may be no employees who drafted or reviewed specific outputs, specific employees may make decisions regarding a tool’s design or monitoring that makes them effectively responsible for individual outputs. *See Sullivan*, 376 U.S. at 287. In particular, courts should be willing to impute actual malice to generative AI developers in at least the following two pre- or post-deployment scenarios where a developer’s conduct suggests a reckless disregard of the risk that a generative AI tool would produce a particular defamatory output.

First, in some cases a developer may make pre-deployment model design decisions that evince a conscious disregard of a substantial and unjustifiable risk of defaming particular public figures. For instance, in 2022 a computer scientist “pranked” the AI research community by training a large language model on content from the internet site 4chan, which he then used to generate 15,000 posts on the site.⁴⁶ 4chan is notorious for hosting extreme and incendiary content, especially about prominent figures including Hillary Clinton and John Podesta.⁴⁷ Since training

⁴⁶ *See* Matthew Gault, *AI Trained on 4chan Becomes ‘Hate Speech Machine,’* Vice (June 7, 2022), <https://www.vice.com/en/article/ai-trained-on-4chan-becomes-hate-speech-machine>.

⁴⁷ *See* Mike Wendling, *The Saga of ‘Pizzagate’: The Fake Story that Shows How Conspiracy Theories Spread*, BBC News (Dec. 2, 2016), <https://www.bbc.com/news/blogs-trending-38156985> (discussing 4chan as the origin of a conspiracy theory that Hillary Clinton was operating a pedophile ring out of a Washington, D.C. pizza restaurant); Andrew Griffin, *What is QAnon? The Origins of Bizarre Conspiracy Theory Spreading Online*, Independent (Jan. 21, 2021), <https://www.independent.co.uk/tech/what-is-qanon-b1790868.html> (identifying 4chan as the

on data from 4chan would make a generative AI model more likely to output similar types of incendiary content, the decision to train a model specifically using 4chan data could support an inference of actual malice towards these public figures of interest, so long as the developer is aware of the widespread existence of untrue and damaging statements about them in the training data.

Second, a failure to screen out future defamatory outputs could amount to actual malice in narrow circumstances. Last year, an Australian mayor threatened to sue OpenAI unless the company prevented ChatGPT from continuing to falsely claim that he had been convicted of bribery.⁴⁸ In response, OpenAI began filtering out all ChatGPT outputs that reiterated the same claim.⁴⁹ But imagine if OpenAI refused to do so, even after this mayor put the company on notice that its generative AI tool was producing these false and harmful outputs. That scenario would come close to the facts of *Harte-Hanks Communications, Inc. v. Connaughton*. In that case, the Supreme Court reaffirmed the principle that “failure to investigate will not alone support a finding of actual malice.” 491 U.S. 657, 692 (1989). However, when a public official had affirmatively proffered recorded tapes which would seriously undermine the story a newspaper intended to publish, and the paper refused to even listen to the tapes, the Court found the paper’s conduct suggested “an intent to avoid the truth” sufficient to infer actual malice. *Id.* at 693. In line with this result, actual malice should be attributed to a generative AI company for repeated defamatory outputs if, despite having the capability to filter out the specific statements at issue, the company was provided credible notice that a particular repeated output by their tool was false and defamatory and took no action in response.

origin of QAnon and a web of conspiracy theories involving Hillary Clinton and figures associated with her like John Podesta or Huma Abedin).

⁴⁸ See Belanger, *supra* note 43.

⁴⁹ *Id.*

As it has historically been in defamation cases, actual malice should be very difficult for a plaintiff to demonstrate against a generative AI company, but it should not be considered impossible—and certainly not simply because that company’s business model ensures that no employee is “aware” of “computer outputs” as they are produced. Def.’s Mem. Supp. Mot. Summ. J. 17, 22 n.3. Although no single generalizable test can be stated, courts would be justified in inferring actual malice whenever generative AI companies take steps that evince a conscious disregard of a substantial risk of harmful and false outputs about particular public figures or officials being produced by their tools. Both examples above suggest a developer’s conscious awareness of some type of risk, either because of a deliberate choice to train on a particular data source to mimic its content and style, or because direct notice creates knowledge of a particular risk. In this case, if the Court determines that Mr. Walters is a public figure, it should focus its decision on the question of whether OpenAI had any knowledge of a particularized risk to Mr. Walters’s reputation.

CONCLUSION

For the foregoing reasons, *amicus* respectfully urges this Court to reject the extreme arguments advanced by the parties in this case, and instead issue a ruling that properly balances the core purpose of defamation law to provide redress for reputational harm with the First Amendment interests in the creation and use of generative AI tools that benefit both speakers and listeners.

Dated: December 12, 2024

Respectfully submitted,

/s/ Clare R. Norins

Clare R. Norins*

Georgia Bar No. 575364

FIRST AMENDMENT CLINIC

University of Georgia School of Law

P.O. Box 388

Athens, Georgia 30603

(706) 542-1419

cnorins@uga.edu

Counsel for Amicus Curiae

**Local counsel for Technology Law and Policy*

Clinic authors Micah Musser, Catherine

Wang, and Jake Karr

**IN THE SUPERIOR COURT OF GWINNETT COUNTY
STATE OF GEORGIA**

MARK WALTERS,

Plaintiff,

v.

OPENAI, L.L.C.,

Defendant.

CIVIL ACTION NO. 23-A-048

**NOTICE OF ELECTRONIC FILING
AND CERTIFICATE OF SERVICE**

I hereby certify that, on this 12th day of December, 2024, I electronically filed the foregoing *Brief of Amicus Curiae Technology Law and Policy Clinic at New York University School of Law in Support of Neither Party* with the Clerk of Court using the efileGA Odyssey system, which will send e-mail notification of such filing to all attorneys of record.

Dated: December 12, 2024

Respectfully submitted,

/s/ Clare R. Norins

Clare R. Norins

Georgia Bar No. 575364