LARGE LIBEL MODELS? LIABILITY FOR AI OUTPUT

Eugene Volokh[*]

Introduction 491
I. The Current Legal Backdrop 494
  A. 47 U.S.C. § 230 494
  B. Reasonably Perceived as Factual Assertions 498
  C. The Inadequacy of Disclaimers About Possible Errors 500
    1. No contractual waiver 500
    2. Still reasonably seen as factual assertions 500
    3. The rumor analogy 501
    4. The quotation-together-with-rebuttal analogy 503
  D. Publication 504
  E. Publication Without a Human Speaker 508
  F. Damages 510
  G. The Pure Economic Loss Rule 511
  H. The Uncopyrightability Argument 512
  I. First Amendment Protection 513
  J. Knowing or Reckless Falsehoods 514
    1. A notice-and-blocking model? 514
    2. The imperfections of notice-and-blocking 518
    3. The bookstore/newsstand/property owner analogy 520

---

[*] Gary T. Schwartz Distinguished Professor of Law, UCLA School of Law (volokh@law.ucla.edu). Thanks to Jane Bambauer, Curtis Bradley, Adam Chilton, Dhammika Dharmapala, John Donahue, Evelyn Douek, David Greene, James Grimmelmann, Genevieve Lakier, Thomas Healy, Haym Hirsh, Saul Levmore, Michael McConnell, Mihai Surdeanu, Anton Troynikov, and Amy Winograd for their helpful comments.

- 4. The single publication rule 521
- K. Negligence 522
  - 1. Responsibility for the equipment a company uses 522
  - 2. The design defect liability analogy 523
  - 3. Possible precautions: Quote-checking 526
  - 4. Possible precautions: Avoiding quotes altogether 529
  - 5. Possible precautions: Double-checking output 530
  - 6. Possible precautions: Notice-and-blocking as to reported falsehoods 531
  - 7. Possible precautions: Discontinuing earlier versions when new versions prove materially more reliable 531
  - 8. Other possible "reasonable alternative design[s]" 532
  - 9. The need for some attention to libel-related risks 532
  - 10. The problem of multiple creators of AI programs 533
- L. The CheckBKG Analogy 536
- M. Aggregate Costs of Liability 539
- II. Should Current Law Be Changed? 540
  - A. Immunity from Negligence Claims? 540
    - 1. Generally 540
    - 2. The first stab privilege 542
    - 3. Specialized AI programs aimed at "learned intermediaries" ... 546
  - B. No Immunity from Claims of Knowing or Reckless Failure to Respond When Put on Notice 547
  - C. An Administrative Agency Solution? 547
- III. Other Torts 548
  - A. Defamation of Corporations 548
  - B. Trade Libel and Slander of Title 549
  - C. False Light 549
  - D. Disclosure of Private Facts 550
  - E. Disclosure of Trade Secrets 551
  - F. Right of Publicity Infringement 551

G. False Statements That Are Likely to Lead to Injury 552
H. Accurate Statements That Are Likely to Facilitate Crime by Some Readers 553
Conclusion 554
Appendix: Other Examples of Fake Defamatory Quotes 555
A. GPT-4 Example 555
B. Bing Example.. 556
C. Google Bard Example 556

INTRODUCTION

In March 2023, I asked ChatGPT-3.5 "What was [R.R.] accused of?" (I used R.R.'s full name, but I'll omit it here, for reasons that will become obvious.) The response began with a somewhat imprecise but mostly accurate statement of what R.R. had indeed been publicly accused of.[1] But then ChatGPT added that,

> In 2019, [R.] pleaded guilty to conspiracy to commit wire fraud and was sentenced to 3 years of probation and ordered to pay a fine.

When I probed further about the guilty plea, ChatGPT referred to a Reuters article, and even gave me a quote:

> Certainly, here is a longer quote from the Reuters article published on December 5, 2018, that mentions R.R.:
>
> "[R.R.], 35, pleaded guilty to one count of conspiracy to commit wire fraud and one count of aggravated identity theft, according to the U.S. Department of Justice. Prosecutors said that [R.], who operated an online reputation management company, and co-conspirators created fraudulent court orders and other legal documents to convince Google to remove negative reviews and other content about his clients. The scheme, which was carried out between 2014 and 2017, was designed to manipulate Google's search rankings in favor of [R.]'s clients, according to prosecutors. [R.] is scheduled to be sentenced on March 8, 2019."

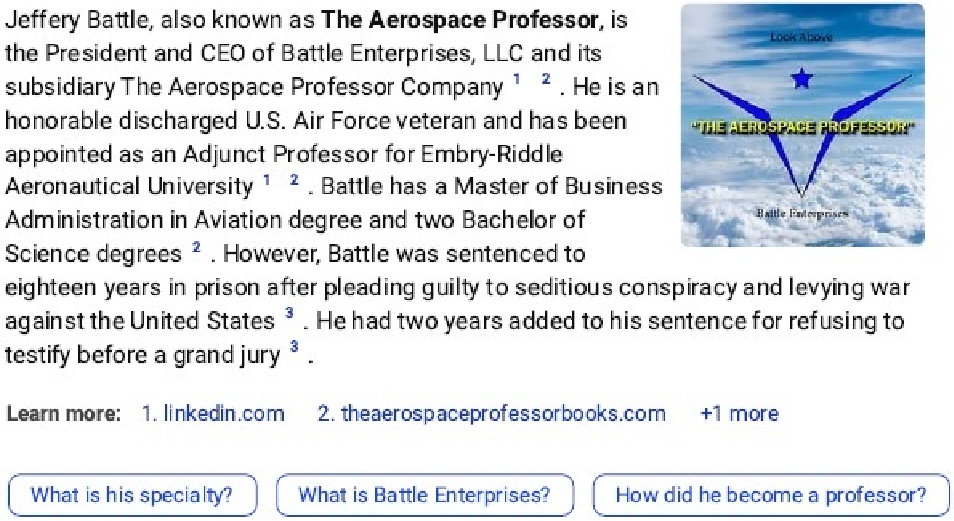
But no such quote seems to appear in any archives of Reuters articles. I could find no record of any guilty plea by R.R. None of the other quotes that ChatGPT-3.5 passed along (in quotation marks) checked out, either. ChatGPT-4 and Bard have

---

[1] He had been publicly accused of this by Paul Alan Levy and me, in Eugene Volokh & Paul Alan Levy, *Dozens of Suspicious Court Cases with Missing Defendants Aim at Getting Web Pages Taken Down or Deindexed*, WASH. POST, Oct. 10, 2016. We did *not* allege that he pleaded guilty.

similarly invented false allegations, complete with invented quotes and invented newspaper articles—I give some examples in the Appendix, though I have more in my files. These Large Language Models (LLMs[2]) seem to routinely erroneously produce false and defamatory statements.[3]

Indeed, a libel lawsuit against OpenAI has already been filed, based on a claim that ChatGPT falsely summarized a complaint in a different case as alleging embezzlement by a particular person; that complaint actually had nothing to do with that person, or with embezzlement.[4] Likewise, a libel lawsuit against Bing has been filed, based on a claim that Bing (which uses GPT-4 technology) responded to a query about "Jeffery Battle" with the following output:



Jeffery Battle, also known as **The Aerospace Professor**, is the President and CEO of Battle Enterprises, LLC and its subsidiary The Aerospace Professor Company [1] [2]. He is an honorable discharged U.S. Air Force veteran and has been appointed as an Adjunct Professor for Embry-Riddle Aeronautical University [1] [2]. Battle has a Master of Business Administration in Aviation degree and two Bachelor of Science degrees [2]. However, Battle was sentenced to eighteen years in prison after pleading guilty to seditious conspiracy and levying war against the United States [3]. He had two years added to his sentence for refusing to testify before a grand jury [3].

Learn more: 1. linkedin.com 2. theaerospaceprofessorbooks.com +1 more

Look Above
"THE AEROSPACE PROFESSOR"
Battle Enterprises
What is his specialty?
What is Battle Enterprises?
How did he become a professor?

This output apparently mixes information about the technology expert Jeff*ery* Battle with information about the convicted terrorist Jeff*rey* Battle, and adds material that conveys the message that the output is all about one person (the word

---

[2] They are called Large Language Models because they are trained on a vast array of documents and gather from them information about the frequency of words appearing near each other.

[3] Others have likewise observed that AI programs' answers have included spurious allegations of plagiarism, hacking, and murder. *See, e.g.*, Ted Rall, *ChatGPT Libeled Me. Can I Sue?*, WALL ST. J., Mar. 16, 2023; Natasha Lomas, *Who's Liable for AI-Generated Lies?*, TECHCRUNCH (June 1, 2022), https://perma.cc/X4PV-ZW3B; Avery Rowe, *Don't Ask an AI for Plant Advice*, TRADESCANTIA HUB (Mar. 12, 2023), https://perma.cc/29TK-XRVG.

[4] Complaint, Walters v. OpenAI, L.L.C., No. 23-A-04860-2 (Ga. Super. Ct. Gwinnett Cty. filed June 5, 2023), *removed to* Walters v. OpenAI, L.L.C., No. 1:23-cv-03122 (N.D. Ga. July 14, 2023).

"However," coupled with the last name "Battle" without inclusion of the slightly different first name).[5] Yet it appears that AI companies have focused little on the risk of libel, though they have focused on many other kinds of risks.

This is becoming especially significant as such AI programs are becoming integrated into search engines (such as Bing) and other products. If people were to end up viewing AI programs as merely fun toys that can't be trusted with regard to any important information—or just as generators of amusing fiction—then such false statements would cause little harm. But, as I'll discuss in Part I.B, I expect that many users will indeed view the AI programs as reliable enough that the users might, for instance, decide not to select one of dozens of job applicants, or not to deal with some professional or service provider, because of an allegation that the program outputs. And even if users realize that AI programs are no more reliable than, say, rumor or gossip, the law generally recognizes that rumor and gossip can be quite damaging, and can therefore be actionable.

Should, then, the AI programs' creators and operators, such as OpenAI (for ChatGPT[6]) or Google (for Bard) be liable for defamation, based on their programs' output?[7] Part I will analyze this question under the current rules of U.S. defamation law. I will tentatively argue that, when the "actual malice" standard applies, the standard might be satisfied if an AI company has received actual notice of particular spurious information being produced by its software but has refused to act. This would in practice require such companies to implement a "notice-and-blocking" system, loosely similar to "notice-and-takedown" systems required under the DMCA as to copyright and trademark infringements. And I will also discuss the possibility of negligence liability, when such liability is authorized under libel law, by analogy to negligent design product liability.

To be sure, allowing such liability could yield substantial costs. That is particularly so since it may require lay judges and juries to evaluate complicated technical

---

[5] Complaint, Battle v. Microsoft Corp. No. 1:23-cv-01822, at 2 (D. Md. filed July 7, 2023). The Bing results are from a query that I ran to verify the allegations in the Complaint, and they closely track the results quoted in the Complaint.

[6] *See* OpenAI, *GPT-4 Technical Report* (Mar. 27, 2023), https://perma.cc/U9VV-M5UA.

[7] I focus here on the possible liability of the AI companies, not of entities (whether individual users or media outlets) that republish material that is produced for them by AIs. *See* Seth C. Lewis, Amy Kristin Sanders & Casey Carmody, *Libel by Algorithm? Automated Journalism and the Threat of Legal Liability*, 96 J. & MASS COMM. Q. 60, 73 (2018); *infra* note 21.

claims about which designs are feasible. (Such concerns of course mirror the concerns about legal liability as to other products, such as pharmaceuticals or cars, or as to services, such as surgical procedures.) Part II will tentatively discuss some arguments for why the law might be changed, whether by courts, by legislatures, or by administrative agencies. Finally, Part III will offer some similarly tentative thoughts about how this might apply to other claims, such as false light, disclosure of private facts, the right of publicity, or negligence.

A terminological note: For convenience, I'll generally use the term "AI programs" to mean AIs that output information in response to a user prompt, though I realize that there are of course AI programs that do other things, such as operate self-driving cars. I'll also refer to "AI companies" as shorthand for companies that create AI programs and provide access to their services.[8] My logic ought to apply to such AI programs however they are designed, whether they are based on LLMs or not, though I appreciate that parts of the analysis may turn on the specifics of each design.

I. THE CURRENT LEGAL BACKDROP

A. 47 U.S.C. § 230

To begin with, 47 U.S.C. § 230 likely doesn't provide AI companies with immunity for material composed and communicated by their AI programs. Section 230 states that, "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by *another* information content provider."[9]

But AI programs' output is composed by the programs themselves—it isn't merely quotations from existing sites (as with snippets of sites offered by search engines[10]) or from existing user queries (as with some forms of autocomplete that recommend the next word or words by essentially quoting them from user-

---

[8] Today, the providers of access to the leading AI programs (GPT and Bard) are also their creators, or the creators' business partners (such as Microsoft, which uses GPT-4 to power Bing). I therefore generally don't distinguish liability for the AI provider from liability for the AI creator. For a few thoughts on liability in such situations, including when the AI provider has adapted the AI creator's program, see *infra* Part I.K.10.

[9] 47 U.S.C. § 230(c)(1) (emphasis added).

[10] *See* O'Kroley v. Fastcase, Inc., 831 F.3d 352, 355 (6th Cir. 2016).

provided content[11]). A lawsuit against an AI company would thus aim to treat it as a publisher or speaker of information provided by *itself*—with it itself being a potentially liable "information content provider," a term that § 230 defines to cover "any person or entity that is *responsible, in whole or in part, for the creation or development of information* provided through the Internet or any other interactive computer service."[12]

As the leading early § 230 precedent, *Zeran v. AOL*, pointed out, in § 230 "Congress made a policy choice . . . not to deter harmful online speech through the . . . route of imposing tort liability on companies that serve as intermediaries for other parties' potentially injurious messages."[13] But Congress didn't make the choice to immunize companies that deploy software which itself creates messages that had never been expressed by third parties.[14]

Likewise, courts have held that § 230 doesn't immunize defendants who "materially contribut[e] to [the] alleged unlawfulness" of online content.[15] An AI company, by making and distributing an AI program that creates false and reputation-damaging accusations out of text that entirely lacks such accusations, is surely

---

[11] For more on whether autocomplete suggestions may be defamatory, see, *e.g.*, Michael L. Smith, *Search Engine Liability for Autocomplete Defamation: Combating the Power of Suggestion*, U. ILL. J.L. TECH. & POL'Y 313; Kacy Popyer, *Cache-22: The Fine Line Between Information and Defamation in Google's Autocomplete Function*, 34 CARDOZO ARTS & ENT. L.J. 835 (2016).

[12] 47 U.S.C. § 230(f)(3) (emphasis added). I thus agree on this with Matt Perault, *Section 230 Won't Protect ChatGPT*, 3 J. FREE SPEECH L. 363 (2023); *see also* Peter Henderson, Tatsunori Hashimoto & Mark Lemley, *Where's the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589, 620–25 (2023).

[13] 129 F.3d 327, 330–31 (4th Cir. 1997).

[14] The statement in *Fair Housing Council of San Fernando Valley v. Roommates.com, LLC*, 521 F.3d 1157, 1175 (9th Cir. 2008) (en banc), that "If you don't encourage illegal content, or design your website to require users to input illegal content, you will be immune," dealt with websites that republish "user[]" "input." The case didn't provide immunity to websites that themselves create illegal (*e.g.*, libelous) content.

[15] *Id.* at 1167–68. Many other courts have endorsed this formulation. FTC v. LeadClick Media, LLC, 838 F.3d 158, 174 (2d Cir. 2016); Jones v. Dirty World Ent. Recordings LLC, 755 F.3d 398, 410 (6th Cir. 2014); FTC v. Accusearch Inc., 570 F.3d 1187, 1200 (10th Cir. 2009); People v. Bollaert, 248 Cal. App. 4th 699, 719 (2016); Vazquez v. Buhl, 150 Conn. App. 117, 135–36 (2014); Hill v. StubHub, Inc., 219 N.C. App. 227, 238 (2012).

"materially contribut[ing] to [the] alleged unlawfulness" of that created material.[16] The program is not a mere "neutral conduit for [the actionable] content"[17]—indeed, it is not a conduit at all.

To be sure, LLMs appear to produce each word (or portion of a word, called a token[18]) based on word frequency connections drawn from sources in the training data. Their output is thus in some measure derivative of material produced by others.[19]

But of course, all of us who are writing our own material rely almost exclusively on words that exist elsewhere, and then arrange them in an order that likewise stems largely from our experience reading material produced by others. Yet that can't justify immunity for us when we assemble others' individual words in defamatory ways.

For instance, courts have read § 230 as protecting even individual human decisions to copy and paste particular material that they got online into their own posts: If I post to my blog some third-party-written text that was intended for use on the Internet (for instance, because it's already been posted online), I'm immune from liability.[20] But of course if I instead myself write a new defamatory post about you, I lack § 230 immunity even if I copied each word from a different web page and then assembled them together: I'm responsible in part (or even in whole) for creating the defamatory information. Likewise for AI programs.

---

[16] If the AI program merely accurately "restat[es] or summariz[es]" material in its training data, even if it doesn't use the literal words, it may still be immune. *See* Derek Bambauer & Mihai Surdeanu, *Authorbots*, 3 J. FREE SPEECH L. 375, 383 (2023). But I'm speaking here of situations where the AI program does "produced . . . new semantic content" rather than "merely repackag[ing] existing content." *Id.*

[17] *Accusearch*, 570 F.3d at 1198–99 ("to be 'responsible' for the development of offensive content, one must be more than a neutral conduit for that content").

[18] *See* Tabarak Khan (OpenAI), *What Are Tokens and How to Count Them?*, https://perma.cc/SWQ7-HXJF.

[19] Jess Miers, *Yes, Section 230 Should Protect ChatGPT and Other Generative AI Tools*, TECHDIRT, Mar. 17, 2023, 11:59 am, https://perma.cc/ZH73-N3XA.

[20] *See, e.g.*, Batzel v. Smith, 333 F.3d 1018, 1026 (9th Cir. 2003), *superseded in part by statute on other grounds as stated in* Breazeale v. Victim Servs., Inc., 878 F.3d 759, 766–67 (9th Cir. 2017); Barrett v. Rosenthal, 146 P.3d 510 (Cal. 2006); Phan v. Pham, 182 Cal. App. 4th 323, 324–28 (2010); Monge v. Univ. of Pennsylvania, No. CV 22-2942, 2023 WL 2471181, *3 (E.D. Pa. Mar. 10, 2023); Novins v. Cannon, No. CIV 09-5354, 2010 WL 1688695, *2 (D.N.J. Apr. 27, 2010).

And this makes sense. If Alan posts something defamatory about Betty on his WordPress blog, that can certainly damage her reputation, especially if the blog comes up on Google searches—but at least people will recognize it as Alan's speech, not Google's or WordPress's, so § 230 treats it just as Alan's speech. But text distributed by an AI company (via its AI program) and framed as the program's own output will be associated in the public's mind with the credibility of the program and the company. That may make it considerably more damaging and would make it fair to hold the company liable for that.

Relatedly, traditional § 230 cases at least in theory allow *someone*—the actual creator of the speech—to be held liable for it (even if in practice the creator may be hard to identify, or outside the jurisdiction, or lack the money to pay damages). Allowing § 230 immunity for libelous output by an AI program would generally completely cut off any recourse for the libeled person, against anyone.[21]

In any event, as noted above, § 230 doesn't protect entities that "materially contribut[e] to [the] alleged unlawfulness" of online content.[22] And when AI programs output defamatory text that they have themselves assembled, word by word, they are certainly materially contributing to its defamatory nature. Consider the ChatGPT output, in the example with which I started the article:

> Certainly, here is a longer quote from the Reuters article published on December 5, 2018, that mentions R.R.:
>
> "[R.R.], 35, pleaded guilty to one count of conspiracy to commit wire fraud and one count of aggravated identity theft, according to the U.S. Department of Justice. Prosecutors said that [R.], who operated an online reputation management company, and co-conspirators created fraudulent court orders and other legal documents to convince Google to remove negative reviews and other content about his clients. The

---

[21] There would generally be no liability for a user who believes an AI program's false output about a person and as a result declines to do business with the person—just as there is generally no liability for someone who believes a false rumor about a person and as a result shuns that person. *See infra* p. 544.

If the user republishes the AI program's output, then the user might be held liable for defamation. This would depend on (1) whether the user is seen as having the requisite culpable *mens rea* (negligence or "actual malice," as the case may be), and on (2) whether the user himself would be immune under § 230(c)(1) on the theory that the user can't be liable for "information provided by another content provider," namely the AI company. But in many situations, the AI program's output would cause harm even if the user doesn't republish it.

[22] *See supra* note 15.

> scheme, which was carried out between 2014 and 2017, was designed to manipulate Google's search rankings in favor of [R.]'s clients, according to prosecutors. [R.] is scheduled to be sentenced on March 8, 2019."

Though each of the words that ChatGPT output presumably existed somewhere in the training data, OpenAI is "responsible, in whole or in part,"[23] for making the quote say what it says, and it "materially contribut[ed] to [the] alleged unlawfulness" (here, the defamatory nature) of the quote.

B. Reasonably Perceived as Factual Assertions

Some have argued that AI programs' output shouldn't be seen as a factual claim, because it's just the result of a predictive algorithm that chooses the next word based on its frequent location next to the neighboring ones in the training data—I've seen analogies to Ouija boards, Boggle, "pulling Scrabble tiles from the bag one at a time," and a "typewriter (with or without an infinite supply of monkeys)."

But I don't think that's right. In libel cases, the threshold "key inquiry is whether the challenged expression, however labeled by defendant, would reasonably appear to state or imply assertions of objective fact."[24] And OpenAI has touted ChatGPT as a generally pretty reliable[25] (though not "fully reliable"[26]) source of assertions of fact, not just as a source of entertaining nonsense. Google's Bard has been less promoted, but presumably it aims to be seen as a worthy rival to ChatGPT.

What is OpenAI doing when it promotes ChatGPT's ability to get high scores on bar exams or the SAT?[27] Or when it stresses in the subtitle to its product description, "GPT-4 can solve difficult problems with greater accuracy, thanks to its broader general knowledge and problem-solving abilities"[28]? It's trying to get the public to view ChatGPT's output as pretty trustworthy. Likewise when its software is incorporated into search engines, or into other applications, presumably

---

[23] 47 U.S.C. § 230(f)(3).

[24] Takieh v. O'Meara, 497 P.3d 1000, 1006 (Ariz. Ct. App. 2021).

[25] *E.g.*, OpenAI, *GPT-4*, https://perma.cc/9UJ5-Y4QG (stressing how successful GPT-4 was on various standardized tests); *id.* (describing GPT-4 as "more reliable . . . than GPT-3.5").

[26] OpenAI, *GPT-4 Technical Report*, *supra* note 6, at 1.

[27] *Id.* at 5; *see also* OpenAI, *GPT-4*, https://perma.cc/9UJ5-Y4QG.

[28] OpenAI, *GPT-4 Is OpenAI's Most Advanced System, Producing Safer and More Useful Responses*, https://perma.cc/GB4P-FPEA.

precisely because it's seen as pretty reliable.[29] The AI companies' current and future business models rest entirely on their programs' credibility for producing reasonably accurate summaries of the facts.

Indeed, as OpenAI noted, "hallucinations"—meaning the inclusion of incorrect information made up by the AI program itself—"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."[30] After OpenAI promotes its superiority to 90% of test-takers at producing answers to complicated questions, it can't then turn around and, in a libel lawsuit, argue that it's all just Jabberwocky.

Naturally, everyone understands that AI programs aren't perfect. But everyone understands that newspapers aren't perfect either, and some are less perfect than others—yet that can't be enough to give newspapers immunity from defamation liability; likewise for AI programs. And that's especially so when the output is framed in quite definite language, complete with purported quotes from respected publications.[31]

To be sure, people who are keenly aware of the Large Libel Models problem might be so skeptical of anything AI programs output that they wouldn't perceive any of the programs' statements as factual. But libel law looks at the "natural and probable effect" of assertions on the "average lay reader,"[32] not at how something is perceived as a technical expert.

---

[29] *See, e.g.*, Yusuf Mehdi, Corporate V.P. & Consumer Chief Marketing Officer, *Microsoft, Reinventing Search with a New AI-Powered Microsoft Bing and Edge*, OFFICIAL MICROSOFT BLOG, Feb. 7, 2023 (describing the GPT-powered Bing as providing "Better search," "Complete answers"); *id.* ("It takes key learnings and advancements from ChatGPT and GPT-3.5—and it is even faster, more accurate and more capable.").

[30] OpenAI, *GPT-4 Technical Report*, *supra* note 6, at 46.

[31] *See also* Henderson, Hashimoto & Lemley, *supra* note 12, at 592 ("People may be inclined to believe [ChatGPT's] statements, for several reasons: (1) human experience with similarly authoritative-seeming stories from the real world suggests that they are generally true, (2) ChatGPT is quite good at accurately reporting facts in many settings, and (3) people don't understand how ChatGPT works or that it suffers from hallucinations.")

[32] *E.g.*, Kelly v. Arrington, 624 So. 2d 546, 548–49 (Ala. 1993) (cleaned up); Fry v. Lee, 408 P.3d 843, 847 (Colo. Ct. App. 2013) (cleaned up); Yancey v. Hamilton, 786 S.W.2d 854, 858–59 (Ky. 1989) (cleaned up); Nelson v. Am. Hometown Publ'g, Inc., 333 P.3d 962, 974 (Okla. Ct. Civ. App. 2014) (cleaned up).

C. The Inadequacy of Disclaimers About Possible Errors

1. No contractual waiver

AI programs often include disclaimers that stress the risk that their output will contain errors. Bard, for instance, includes under the prompt box, in small print, "Bard may display inaccurate or offensive information that doesn't represent Google's views." ChatGPT likewise notes, "ChatGPT may produce inaccurate information about people, places, or facts." But such disclaimers don't immunize AI companies against potential libel liability.

To begin with, such disclaimers can't operate as contractual waivers of liability: Even if the AIs' *users* are seen as waiving their rights to sue based on erroneous information when they submit a query despite seeing the disclaimers, that can't waive the rights of the *third parties* who might be libeled.

2. Still reasonably seen as factual assertions

Nor do the disclaimers keep the statements from being reasonably viewed as actionable false statements of fact. Defamation law has long treated false, potentially reputation-damaging assertions about people as actionable even when it's evident that the assertions might be false. No newspaper can immunize itself from libel lawsuits for a statement that "Our research reveals that John Smith is a child molester" by simply adding "though be warned that this might be inaccurate" (much less by putting a line on the front page, "Warning: We may sometimes publish inaccurate information"). Likewise, if I write "I may be misremembering, but I recall that Mary Johnson had been convicted of embezzlement," that could be libelous despite my "I may be misremembering" disclaimer.

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. In libel cases, a "key inquiry is whether the challenged expression, however labeled by defendant, would reasonably appear to state or imply assertions of objective fact."[33] It's not actionable to state something that obviously contains no factual

---

[33] Takieh v. O'Meara, 497 P.3d 1000, 1006 (Ariz. Ct. App. 2021).

assertion at all.[34] But it is actionable to state a factual assertion about which the speaker expresses uncertainty.

And neither ChatGPT nor Bard actually describe themselves as generally producing fiction (at least in the absence of a prompt that asks them to produce fiction), since that would be a poor business model for them. Rather, they tout their general reliability,[35] and simply acknowledge the risk of error. That acknowledgment doesn't preclude liability.[36]

3. The rumor analogy

The law's treatment of rumors offers a helpful analogy. When reasonable people hear the statement, "Rumor has it that John Smith was convicted of embezzlement," they recognize that the underlying assertion about the conviction might be true but might be false. "Rumor has it" is in practice a form of disclaimer, since all of us are aware that rumors are often untrustworthy.

Yet "when a person repeats a slanderous charge, even though identifying the source or indicating it is merely a rumor, this constitutes republication and has the same effect as the original publication of the slander."[37] When speakers identify something as rumor, they are implicitly saying "this may be inaccurate"—but that doesn't itself get them off the hook.[38] (Discussing rumors in the course of calling for an investigation of the matter, or of reporting on the controversy created by the

---

[34] *See, e.g.*, Greene v. Paramount Pictures Corp., 813 F. App'x 728, 731–32 (2d Cir. 2020). Even then, a court might allow liability if it concludes that a reasonable person who knows plaintiff would understand that defendant's ostensible fiction is actually meant to be a *roman à clef* that conveys factual statements about plaintiff. The presence of a disclaimer wouldn't be dispositive then. *See, e.g.*, Pierre v. Griffin, No. 20-CV-1173-PB, 2021 WL 4477764, *6 n.10 (D.N.H. Sept. 30, 2021).

[35] *See supra* Part I.B.

[36] *Cf.* Rogers v. Mroz, 502 P.3d 986, 989 (Ariz. 2022) ("[Political c]andidates cannot make defamatory assertions they hope voters will believe, then, when sued for defamation, seek refuge in the defense that no one believes what politicians say.").

[37] Ringler Assocs. Inc. v. Maryland Cas. Co., 80 Cal. App. 4th 1165, 1180 (2000).

[38] OpenAI has argued that "there is near universal consensus that responsible use of AI includes fact-checking prompted outputs before using or sharing them." Motion to Dismiss, Walters v. OpenAI, L.L.C., No. 1:23-cv-03122, at 2 (N.D. Ga. July 21, 2023). But of course there is near universal—and much longer-standing—consensus that responsible response to rumors includes fact-checking them before using or sharing them; and yet people may be held liable for their statements even if they label them as rumors.

rumors, might sometimes be seen as not actionable;[39] but that context is absent in the typical AI program output, which usually just frames its statements as factual assertions, even if not completely reliable assertions.)

Indeed, according to the Restatement (Second) of Torts, "the republisher of either a libel or a slander [is] subject to liability even though he expressly states that he does not believe the statement that he repeats to be true."[40] Some recent cases depart from this, and make clear that such an expression of disbelief would keep the statement from being a defamatory factual assertion.[41] But liability can't be prevented by a disclaimer that the statement merely *may* be inaccurate.

There is a narrow rumor privilege that allows a person to repeat certain kinds of rumors to particular people when "the relation of the parties, the importance of the interests affected and the harm likely to be done make the publication reasonable."[42] (This stems from what is seen as the special legitimacy of people protecting friends' interests, or warning people about harm that others may cause them.) This is why, for instance, if Alan tells Betty that he had heard a rumor that Betty's employee Charlie was a thief, Alan is immune from liability.[43] But the privilege exists

---

[39] *See, e.g.*, Brian v. Richardson, 87 N.Y.2d 46, 53 (1995) ("[T]he purpose of defendant's article was to advocate an independent governmental investigation . . . . To support this argument, defendant marshalled the relevant rumors and accusations that were floating around . . . . Indeed, without a recitation of the existing unresolved charges, defendant's call for a full-scale investigation would have made no sense. Given this contextual background, we conclude on this record that a reasonable reader would understand the statements defendant made about plaintiff as mere *allegations* to be investigated rather than as *facts.*"); Vengroff v. Coyle, 231 A.D.2d 624, 625 (N.Y. App. Div. 1996) (following *Brian*); Croce v. N.Y. Times Co., 930 F.3d 787, 794–95 (6th Cir. 2019) ("[T]he article does not say that Dr. Croce is guilty of any of these allegations and charges of scientific misconduct, nor does the article suggest that these allegations are true. If the article suggested that these allegations were true, or if the article did not use language that qualified the statements made by others as allegations, then the Defendants potentially could be liable for reporting third-party statements. In this case, the accusations are couched in terms like, 'Some scientists argue,' 'allegations,' 'claims of,' 'criticisms,' 'charges,' and 'complaints.' . . . In its full and proper context, . . . the article reports newsworthy allegations with appropriate qualifying language.").

[40] RESTATEMENT (SECOND) OF TORTS § 578 cmt. e (1977); *see also* Martin v. Wilson Pub. Co., 497 A.2d 322, 327 (R.I. 1985); Hart v. Bennet, 267 Wis. 2d 919, 944 (App. 2003).

[41] *See, e.g.*, Blesedell v. Chillicothe Tel. Co., 811 F.3d 211, 225 (6th Cir. 2016).

[42] RESTATEMENT (SECOND) OF TORTS § 602 (1977).

[43] *Id.* ill. 2. Another classic illustration is a parent warning an adult child about a rumor that the child's prospective spouse or lover is untrustworthy. *Id.* ill. 1.

precisely because, without it, passing along factual allegations about most topics to most audiences—even with an acknowledgment that they "may [be] inaccurate"—may be actionable.[44]

4. The quotation-together-with-rebuttal analogy

Likewise, say that you present both an accusation and the denial of the accusation. By doing that, you're making clear that the accusation may well be inaccurate—perhaps the accusation is wrong and the response is right. Yet that doesn't stop you from being liable for repeating the accusation.[45]

To be sure, there are some narrow and specific privileges that defamation law has developed to free people to repeat possibly erroneous content without risk of liability, in particular contexts where such repetition is seen as especially necessary. For instance, some courts recognize the "neutral reportage" privilege, which immunizes "accurate and disinterested" reporting of "serious charges" made by "a responsible, prominent organization" "against a public figure," even when the reporter has serious doubts about the accuracy of the charges.[46] But other courts reject the privilege.[47]

And even those that accept it apply it only to narrow situations: Reporting false allegations remains actionable—even though the report makes clear that the allegations may be mistaken—when the allegations relate to matters of private concern, or are made by people or entities who aren't "responsible" and

---

[44] *See, e.g.*, Martin v. Wilson Pub. Co., 497 A.2d 322, 327 (R.I. 1985).

[45] *See, e.g.*, Norton v. Glenn, 860 A.2d 48, 50, 58 (Pa. 2004).

[46] Edwards v. Nat'l Audubon Soc'y, 556 F.2d 113 (2d Cir. 1977). A few later cases have extended this to certain charges on matters of public concern against private figures. Others have rejected the privilege as to statements about private figures, without opining on its availability as to public figures. *See, e.g.*, Khawar v. Globe Int'l, Inc., 965 P.2d 696, 707 (Cal. 1998); Fogus v. Cap. Cities Media, Inc., 444 N.E.2d 1100, 1102 (Ill. Ct. App. 1982).

[47] Norton v. Glenn, 860 A.2d 48, 50, 58 (Pa. 2004); Dickey v. CBS, Inc., 583 F.2d 1221, 1225–26 (3d Cir. 1978); McCall v. Courier-J. & Louisville Times, 623 S.W.2d 882 (Ky. 1981); Postill v. Booth Newspapers, Inc., 325 N.W.2d 511 (Mich. App. 1982); Hogan v. Herald Co., 84 A.D.2d 470, 479 (N.Y. App. Div. 1982).

"prominent."[48] Such reporting certainly remains actionable when the allegations themselves are erroneously recalled or reported by the speaker.

The privilege is seen as needed precisely because of the general rule that, absent such a privilege, passing on allegations can be libelous even when it's made clear that the allegations may be erroneous. And the privilege is a narrow exception justified by the "fundamental principle" that, "when a responsible, prominent organization . . . makes serious charges against a public figure," the media must be able to engage in "accurate and disinterested reporting of those charges," because the very fact that "they were made" makes them "newsworthy."[49]

D. Publication

I've also heard the argument that statements by AIs in response to user queries aren't really "published," because they are just one-to-one responses. But defamation law has always applied to one-to-one writings (such as personal letters,[50] or notes with comments on an ex-employee's job record[51]) and one-to-one oral statements (for instance, in telephone calls[52]). The famous Supreme Court libel cases are mostly about statements in newspapers or magazines; but libel law has never been limited to that. The Restatement (Second) of Torts captures it well, making it clear that "publication" in libel cases is a legal term of art:

---

[48] A few authorities have applied this privilege to accurate reporting of allegations on matters of public concern generally, but this appears to be a small minority. Barry v. Time, Inc., 584 F. Supp. 1110 (N.D. Cal. 1984); TEX. CIV. CODE § 73.005.

[49] *Edwards*, 556 F.2d at 120. Likewise, the fair report privilege allows one to accurately repeat allegations that were made in government proceedings or in "meeting[s] open to the public," because of the deeply rooted principle that the public must be able to know what was said in those proceedings, even when those statements damage reputation. But this privilege is limited to *accurate* repetition of allegations originally made in government proceedings or in meetings open to the public. RESTATEMENT (SECOND) OF TORTS § 611 (1977).

[50] *See, e.g.*, RESTATEMENT (SECOND) OF TORTS § 577 ill. 6 (1977).

[51] Such job references are sometimes qualifiedly privileged, and thus not actionable when "based upon credible evidence [and] made without malice," *see, e.g.*, CAL. CIV. CODE § 47(c); but this privileges is needed precisely because this sort of one-to-one communication would otherwise be libelous.

[52] *See, e.g.*, RESTATEMENT (SECOND) OF TORTS § 577 ill. 7 (1977).

> Publication of defamatory matter is its communication intentionally or by a negligent act to one other than the person defamed.[53]

And the Restatement offers several illustrations of one-to-one communications being libelous.[54] A statement output to a user by an AI program, which is created and operated by an AI company, is thus publication to the user.[55]

For an analogy, consider *Finlay v. MyLife.com*.[56] MyLife offered background reports to online users, in what plaintiff alleged to be a highly automated—and inaccurate—way.[57] The court concluded that plaintiff had "sufficiently alleged a third-party publication" by MyLife because "the Complaint's allegations give rise to a reasonable inference of third-party disclosure."[58] Just as MyLife's conveying false information about Finlay to its users is a publication, so is an AI program's conveying false information about third parties to its users. Many other cases likewise treat communication by a computerized system—such as one run by a credit rating agency—to a user of that system as a publication for libel law purposes.[59]

---

[53] *Id.* § 577(1). A statement said just to the plaintiff—*e.g.*, accusing someone of being a thief, when no-one else is present—can't be libelous because it can't damage the plaintiff's reputation with third parties. Likewise, a person's research notes, read only by that person, don't constitute publication, again because the notes don't damage the plaintiff's reputation: They reflect only what the author of the notes already thought before writing the notes down.

Note that the "intentionally or by a negligent act" in this section refers to the act of *communication*; the formulation precludes liability when, say, a person's note in his desk is unexpectedly seen by a third party (compare *id.* ill. 5, which imposes liability when the note is negligently left where it can be seen). It doesn't refer to knowledge or negligence as to the *falsehood of the statement*; that is the subject of the rules described in Parts I.H–I.J.

[54] *See, e.g.*, RESTATEMENT (SECOND) OF TORTS § 577 ills. 6–9 (1977).

[55] *See also* Henderson, Hashimoto & Lemley, *supra* note 12, at 635 (reaching the same conclusion).

[56] 525 F. Supp. 3d 969 (D. Minn. 2021).

[57] *See id.* at 975 (citations omitted) (alleging that MyLife "searches public records databases, among other sources" and "generates [a] MyLife profile using all of the information it aggregates").

[58] *Id.* at 983.

[59] *See also, e.g.*, Unlimited Cellular, Inc. v. Red Points Solutions SL, __ F. Supp. 3d __, 2023 WL 4029824, *3–*4 (S.D.N.Y. 2023) (finding that plaintiff adequately pleaded publication by alleging that defendants "submitted [their] reports to eBay" by having their "software automatically report[] detected intellectual property infringements to e-commerce websites"); Loomis v. U.S. Bank Home Mortg., 912 F. Supp. 2d 848, 860 (D. Ariz. 2012) (allowing a defamation case to go forward on the

Communications by a program that was custom-made for a company to that company's employees might not constitute a publication. The only case I know of on that topic is *Murray v. ILG Technologies LLC*, which involved a software developer (ILG) that had a contract with the Georgia Office of Bar Admissions "to create and provide a computer program that would facilitate the entire bar admissions process."[60] The program erroneously communicated to the Office that particular people didn't pass the Georgia Bar, but the court held that this wasn't a publication for defamation law purposes:

> To the extent the Bar Applicants argue the transmission of the incorrect bar results from the software to the OBA is itself a "publication," the Bar Applicants have cited no law that supports this contention. And even assuming such a transmission could constitute a publication, Georgia law recognizes "an exception to the broad definition of publication." "[W]hen the communication is intracorporate, or between members of unincorporated groups or associations, and is heard by one who, because of his/her duty or authority has reason to receive [the] information, there is no publication of the allegedly slanderous material . . . ." Under this doctrine, the software's communications to employees of the OBA—to the extent such communications constitute "publication"—would be "intracorporate" and, therefore, nonactionable.[61]

But the intracorporate communications exception is a minority view among states.[62] And even those states that recognize it limit it to situations where the speaker and the listener are both "agent[s]" of the same corporation.[63] OpenAI,

---

theory that a bank "automatically reported to the credit reporting agencies that Plaintiffs' mortgage was in modified status simply because Plaintiffs inquired about the possibility of a mortgage modification"); Shaunfield v. Experian Info. Sols., Inc., 991 F. Supp. 2d 786, 803 (N.D. Tex. 2014) (allowing libel case against a credit reporting agency to go forward, on the theory that the agency had "published false credit information about him to third parties").

[60] 798 F. App'x 486, 488 (11th Cir. 2020).

[61] *Id.* at 494 (cleaned up).

[62] *See* RODNEY A. SMOLLA, LAW OF DEFAMATION § 15:9 (2022); *see, e.g.*, Dube v. Likins, 216 Ariz. 406, 418 (Ct. App. 2007) (taking the view that an intracorporate communication does constitute publication); *see also* RESTATEMENT (SECOND) OF TORTS § 577 cmt. i (1977) (likewise).

[63] *See, e.g.*, Kitchen Hardware, Ltd. v. Kuehne & Nagel, Inc., 421 S.E.2d 550, 553 (Ga. Ct. App. 1992) ("a communication made by one corporate agent to another is not publication in the legal sense") (cleaned up); Brunet v. Fullmer, 777 So. 2d 1240, 1242 (La. Ct. App. 2001) (likewise); Block v. Pre-Paid Legal Servs., Inc., No. CIV-07-1304-F, 2010 WL 11509211, *7 (W.D. Okla. Feb. 4, 2010) (concluding that communications weren't "intra-corporate" unless there was "an agency relation-

Google Bard, and Bing are not agents of its users under agency law principles,[64] just as MyLife and credit rating agencies are not agents of their users. The output of a Bing query to someone searching for "Jeffery Battle," in the example given in the Introduction, is not an "intracorporate" communication.

Some other legal rules require something more like the lay meaning of "publication." For instance, the false light and disclosure of private facts torts have a "publicity" element, which means they apply only when "matter is made public, by communicating it to the public at large, or to so many persons that the matter must be regarded as substantially certain to become one of public knowledge."[65] Likewise, certain copyright law principles turn on whether defendant engaged in "publication," meaning "distribution . . . to the public," or performed or displayed a work "publicly," meaning (among other things) "at a place open to the public or at any place where a substantial number of persons outside of a normal circle of a family and its social acquaintances is gathered."[66] But such publication in the colloquial sense is *not* required for libel liability.[67]

Of course, even if publication to a substantial group of people were required (as would be the case for the false light tort, see Part III.A), that could still be found when a statement, even with some variation, was distributed to many people at different times. Indeed, the copyright law definition of what counts as "public[]" performance of a copyrighted work (such as a song) recognizes that:

---

ship between the corporation" that employed the recipient of the communication "and the independent contractor[]" that conveyed the communication); Neal v. City of Bainbridge Island, No. 3:20-CV-06025-DGE, 2023 WL 3004187 (W.D. Wash. Apr. 19, 2023) ("'For a corporation . . . acting through one of its agents or representatives, to send a libelous communication to 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.'" (quoting Washington state precedent)).

[64] *Cf.* RESTATEMENT (THIRD) OF AGENCY § 1.01 & cmt. e (2006) (defining agency as a "fiduciary relationship"). Simply using a third party's off-the-shelf software doesn't make the third party the user's agent in the legal sense. For whatever it's worth, OpenAI's Terms of Use specify that they "do not create . . . [an] agency relationship between you and OpenAI." OpenAI, *Terms of Use* (Mar. 14, 2023), https://perma.cc/6DS3-CQ9V.

[65] *See* RESTATEMENT (SECOND) OF TORTS §§ 652D cmt. a, 652E cmt. a (1977).

[66] 17 U.S.C. § 101.

[67] *See* RESTATEMENT (SECOND) OF TORTS § 652D cmt. a (1977) (reaffirming that the "publication" element of the libel tort, unlike the "publicity" element of the false light and disclosure of private facts torts, "includes any communication by the defendant to a third person").

> To perform or display a work "publicly" means—
>
> (1) to perform or display it at a place open to the public or at any place where a substantial number of persons outside of a normal circle of a family and its social acquaintances is gathered; or
>
> (2) to transmit or otherwise communicate a performance or display of the work to a place specified by clause (1) or to the public, by means of any device or process, *whether the members of the public capable of receiving the performance or display receive it in the same place or in separate places and at the same time or at different times.*[68]

And this makes sense. After all, if I post something on my web site, it will only be communicated to readers one at a time as they visit it, perhaps one today, one next week, another the week after, and so on. Yet that would still constitute giving "publicity" to the information for false light or disclosure of private facts purposes.[69] Likewise if an AI program communicates information about a person on separate occasions to different readers.

E. Publication Without a Human Speaker

Some have also argued that "publication" only covers communication where the particular words were deliberately crafted or at least selected by a human being (rather than by an algorithm). But I don't think that's right: Errors in what a company communicates can be defamatory regardless of whether the errors stem from direct human error in composing text or from human error in creating the technology that produces the text.[70]

Consider an example, based on a libel case that arose out of the Whitewater investigations during the Clinton Administration: The *Arkansas Democrat-Gazette* wrote a story about the indictment of Arkansas lawyer Eugene Fitzhugh, but included a photograph of a different Arkansas lawyer, J. Michael Fitzhugh. The Arkansas Supreme Court upheld a verdict in favor of J. Michael Fitzhugh, based on

---

[68] 17 U.S.C. § 101. The provision deals with communication of a particular performance, while AI programs output will often have some degree of random variation; but such variation shouldn't make a difference for libel and false light purposes, so long as the same underlying assertion is being communicated.

[69] *See, e.g.*, Safex Found., Inc. v. Safeth, Ltd., 531 F. Supp. 3d 285, 302 & n.8 (D.D.C. 2021).

[70] *See also* Henderson, Hashimoto & Lemley, *supra* note 12, at 636.

the theory that the newspaper was negligent in including the photograph.[71] (A similar fact pattern has also arisen in other cases.[72])

Let's say that the same situation arose today, but the newspaper had created and was using a photograph-retrieval program—AI or otherwise—that would try to find a photograph that matched the story; and say that the program had a bug that sometimes led it to find the wrong person's photograph. Surely the newspaper's display of the wrong photograph would still be a "publication" of the photograph and thus of the defamatory implicit assertion that the person in the photograph is the convicted criminal. In both the real case and the hypothetical, the newspaper would be causing the communication of erroneous, reputation-damaging information, even if it were to do so without any employee consciously focusing on the error.

To be sure, in both cases the plaintiff would generally have to prove negligence (more on that below). But negligence in manually selecting a photograph is comparable here to negligence in designing photograph-finding software, or negligence in continuing to use the software once the newspaper is aware of the bugs. The newspaper would thus be liable for reputation-damaging falsehoods communicated in the newspaper articles that it communicates to readers. Likewise, AI companies can be liable for reputation-damaging falsehoods within the information that their programs communicate to readers—assuming, of course that the AI company has the relevant culpable mental state.[73]

---

[71] *See* Little Rock Newspapers, Inc. v. Fitzhugh, 954 S.W.2d 914, 926 (Ark. 1997). For a similar case, though involving a typo rather than the wrong photograph, see *S. Bell Tel. & Tel. Co. v. Coastal Transmission Serv., Inc.*, 307 S.E.2d 83 (Ga. Ct. App. 1983), where a phone company that produced a Yellow Pages was held liable for misprinting an auto transmission shop's motto "Get it in gear" as "Get it in rear." For an example of a libel-by-typo claim that was rejected on the grounds that the communication was privileged, under the particular facts of the case, see *Whittington v. McGraw-Hill, Inc.*, 294 So. 2d 288 (La. Ct. App. 1974), where a newsletter typist mistyped the company name "Whittington-Banderies Real Estate" as "Shittington-Banderies Real Estate."

[72] *See, e.g.*, *Little Rock Newspapers*, 954 S.W.2d at 918–20 (offering some other examples).

[73] *Cf.* Tomkiewicz v. Detroit News, Inc., 246 Mich. App. 662, 676 (2001) (rejecting liability for publication of the wrong person's photograph on the grounds that the person, a police officer, was a public figure and therefore had to show knowing or reckless falsehood); Peterson v. New York Times Co., 106 F. Supp. 2d 1227, 1230 (D. Utah 2000) (similar); Jones v. New Haven Register, Inc., No. 393657, 2000 WL 157704 (Conn. Super. Ct. Jan. 31, 2000) (similar).

F. Damages

Most state courts view written defamatory publications as actionable even without a showing of "special harm"—*i.e.*, provable economic loss.[74] The First Amendment limits this so-called "presumed damages" doctrine in private figure/public concern cases that are premised on a showing of mere negligent falsehood (as opposed to reckless or knowing falsehood): In such cases, some showing of damage to reputation, and consequent economic loss or emotional distress, is required.[75] But in cases brought based on speech on matters of private concern, or in cases where reckless or knowing falsehood is shown (see Part I.I), damages need not be shown.

In any event, though, damages often could be shown. The results of one response to one user's prompt will likely cause at most limited damage to the subject, and might thus not be worth suing over (though in some situations the damage might be substantial, for instance if the user is deciding whether to hire the subject, or do business with the subject). But of course what one person asks, others might as well; and a subpoena to the AI company, seeking information from any search history logs that the company may keep for its users (as OpenAI and Google do as to ChatGPT and Bard), may well uncover more examples of such queries. And as these AIs are worked into search engines and other products, it becomes much likelier that lots of people will see the same false and reputation-damaging information.

But beyond this, libel law has long recognized that a false and defamatory statement to one person will often be foreseeably repeated to others—and the initial speaker could be held liable for harm that is thus proximately caused by such republication.[76] In deciding whether such repetition is foreseeable, the Restatement

---

[74] RESTATEMENT (SECOND) OF TORTS §§ 569, 575 cmt. b.

[75] Gertz v. Robert Welch, Inc., 418 U.S. 323, 349 (1974). This is the reason that the recently filed AI libel case, Walters v. OpenAI, L.L.C., No. 1:23-cv-03122 (N.D. Ga. removed July 14, 2023), is unlikely to prevail: The statement there is on a matter of public concern (since it involves an assertion about a lawsuit), and there is no claim that OpenAI was informed about the false assertions and nonetheless continued to publish them.

[76] RESTATEMENT (SECOND) OF TORTS § 576(c) (1977); *see, e.g.*, Oparaugo v. Watts, 884 A.2d 63, 73 (D.C. 2005) ("The original publisher of a defamatory statement may be liable for republication if the republication is reasonably foreseeable."); Schneider v. United Airlines, Inc., 208 Cal. App. 3d 71, 75 (1989) (likewise); Brown v. First National Bank of Mason City, 193 N.W.2d 547, 555 (Iowa

tells us, "the known tendency of human beings to repeat discreditable statements about their neighbors is a factor to be considered."[77] Moreover, if the statement lacks any indication that the information should "go no further," that lack "may be taken into account in determining whether there were grounds to expect the further dissemination."[78]

G. The Pure Economic Loss Rule

The prospect of liability in such cases may surprise people who are used to negligence claims against software developers being relatively rare, compared to similar negligence or design defect claims against tangible product manufacturers.[79] Negligence law and design defect law generally applies to behavior that causes *physical injury to people or property*;[80] though some software can proximately cause physical injury, most generally doesn't. Negligence law and design defect law therefore generally excludes *pure economic loss*,[81] so if a bug in a computer program causes computer systems to fail and businesses to lose money, that is generally not the basis for a tort claim.[82]

1972) (likewise). This appears to be the majority view, though some states seem to reject this theory, *see, e.g.*, Fashion Boutique of Short Hills, Inc. v. Fendi USA, Inc., 314 F.3d 48, 60 (2d Cir. 2002).

[77] RESTATEMENT (SECOND) OF TORTS § 576(c) cmt. d (1977).

[78] *Id.*

[79] *See infra* Part I.J.2.

[80] *See, e.g.*, RESTATEMENT (THIRD) OF TORTS: PHYS. & EMOT. HARM §§ 4, 7 (2010).

[81] *See, e.g.*, RESTATEMENT (THIRD) OF TORTS: LIAB. FOR ECON. HARM § 1(1) (2020); RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 21 & cmt. a (1998).

[82] *See, e.g.*, Aspen Am. Ins. Co. v. Blackbaud, Inc., No. 3:22-CV-44 JD, 2022 WL 3868102, *14 (N.D. Ind. Aug. 30, 2022); Opperman v. Path, Inc., 87 F. Supp. 3d 1018, 1054–55 (N.D. Cal. 2014); Shema Kolainu-Hear Our Voices v. ProviderSoft, LLC, 832 F. Supp. 2d 194, 206 (E.D.N.Y. 2010); Gus' Catering, Inc. v. Menusoft Sys., 171 Vt. 556, 558 (2000). *Cf. In re* Fort Totten Metrorail Cases Arising Out of Events of June 22, 2009, 895 F. Supp. 2d 48, 87–88 (D.D.C. 2012) (allowing a products liability claim to go forward based on alleged design defect in software that caused physical harm rather than pure economic loss).

Software bugs also tend to cause economic loss to the companies that are using the software, and those users often waive any right to sue as a condition of the license agreements for the software. Defamation, on the other hand, causes harm to third parties—not, in the situations discussed in this article, to an AI program's users, but to the people whom the AI program mentions in its output to users. Those third parties wouldn't have had occasion to waive any claims against the AI company.

But defamation—including defamation in one-to-one publications and not just in mass media sources[83]—has always been an exception from the pure economic loss rule:[84] Negligent damage to reputation can indeed be the basis for a tort claim.[85] If your software causes a system crash that keeps me from getting a profitable business deal, then you generally won't be liable. But if your software outputs false and defamatory statements about me that keep me from getting a profitable business deal, then you might well be liable.[86]

H. The Uncopyrightability Argument

I've heard some argue that, because the unmodified output of AI programs is generally not copyrightable by the AI company or the user,[87] it shouldn't lead to liability for the AI company, either. But this is a false symmetry. Copyright law and defamation law serve sharply different purposes: Copyright law is supposed to give an incentive to creative expression (without allowing anyone to monopolize facts or ideas), while defamation law is supposed to deter and punish factual assertions that damage someone's reputation.

As a result, it's routine for uncopyrightable statements to be defamatory. If Don, for instance, says "Paula is a convicted embezzler," that isn't copyrightable, because it's an insufficiently expressive short phrase.[88] But may well be defamatory.

Likewise, if a document says "Paula was found not guilty of embezzlement," and Don quotes it while omitting the "not," that isn't copyrightable by Don, since Don hasn't added sufficient expression to the statement. Indeed, if it's a

---

[83] *See infra* Part I.D.

[84] "Wrongs that might seem to cause only economic loss are sometimes regarded otherwise because the law takes an expansive view of what counts as a personal injury. Defamation, for example, is regarded as inflicting a kind of personal injury: harm to the plaintiff's reputation." RESTATEMENT (THIRD) OF TORTS: LIAB. FOR ECON. HARM § 2 cmt. a (2020).

[85] *See infra* Part I.J.

[86] *Cf.* Harter v. RealPage, Inc., 218 F. Supp. 3d 535, 538–40 (E.D. Tex. 2016) (considering defamation claim based on software output, but concluding that the software manufacturer wasn't liable because it accurately communicated public record information and because that particular communication was in any event privileged and made without knowledge or recklessness as to falsehood).

[87] *See* Second Request for Reconsideration for Refusal to Register a Recent Entrance to Paradise (U.S. Copyright Rev. Bd. Feb. 14, 2022), https://perma.cc/7TFE-LA2T.

[88] 37 C.F.R. § 202.1(a).

government document, it might not be copyrightable by anyone.[89] But the altered statement ("Paula was found guilty . . .") may well be defamatory.[90]

The same applies to statements generated by AI programs. Even if the statements are not protected by copyright, that doesn't bear on whether the AI companies are responsible for any defamatory material in the statements.

I. First Amendment Protection

AI programs' output should be as protected by the First Amendment as the output of the *New York Times*, for reasons Mark Lemley, Peter Henderson, and I discuss elsewhere in this issue.[91] Nonetheless, even if an AI program's output is like a newspaper's output (and equally protected by the First Amendment), the AI company would still be potentially exposed to libel liability:

1. If the company knows certain statements the program is communicating are false and defamatory (or if it knows they are likely to be so but recklessly disregards that possibility), then the company could be liable.[92]
2. If the program communicates something false and defamatory about a private figure on a matter of public concern, and the company is negligent about this, then it could be liable for proven harm to the private figure.[93]

---

[89] Documents created by the federal government are uncopyrightable. *See* 17 U.S.C. § 105. So are state court opinions and statutes. *See* Georgia v. Public.Resource.Org, Inc., 140 S. Ct. 1498 (2020); Banks v. Manchester, 128 U.S. 244 (1888).

[90] Even if the original statement was a government document, the altered republication wouldn't be protected by the fair report privilege, because of the alteration. *See* RESTATEMENT (SECOND) OF TORTS § 611 (1977).

[91] *See* Eugene Volokh, Mark A. Lemley & Peter Henderson, *Freedom of Speech and AI Output*, 3 J. FREE SPEECH L. 653 (2023).

[92] N.Y. Times Co. v. Sullivan, 376 U.S. 254 (1964); Curtis Publ'g Co. v. Butts, 388 U.S. 130 (1967). Dallin Albright, *Do Androids Defame with Actual Malice? Libel in the World of Automated Journalism*, 75 FED. COMM. L.J. 103, 115–16 (2022), argues that defamation cases based on AI output should always apply the negligence standard, even for speech about public figures or public officials, in part because "algorithms are designed to produce information mechanically, and it would be impossible to prove they possessed ill will or doubts in the traditional sense." But the question, it seems to me, is whether the *AI companies* know that a statement that the algorithms are communicating is false or likely to be false, not whether the *algorithms* "possess[] . . . doubts."

[93] Gertz v. Robert Welch, Inc., 418 U.S. 323, 347 (1974).

3. If the program communicates something on a matter of private concern, then the company could potentially be strictly liable,[94] though practically speaking almost all states require a showing of negligence even in private-concern cases.[95]

J. Knowing or Reckless Falsehoods

1. A notice-and-blocking model?

Let us consider, then, category 1 above: Material that an AI program communicates that the AI company knows is false (or as to which it recklessly disregards the possibility of falsehood).

It's highly unlikely that the company will know, at the design stage, that the program will be communicating particular defamatory falsehoods about particular people. But say that a person alerts the company that its program is making false assertions about him, and points out that the quotes that its program is reporting as supporting those assertion don't actually appear in the publications—a Lexis/Nexis search and a Google search should verify that—and that there's no record of any federal prosecution of him.[96] Or consider the Jeffery Battle case discussed in the Introduction, where Bing apparently attributed to Jeffery Battle the serious crimes committed by the similarly named Jeffrey Battle, and continued to do so after Jeffery Battle informed Microsoft of the problem.[97]

Someone at the company would then be aware that the company's program is communicating false and defamatory materials.[98] Presumably the company could then add code that would prevent these particular allegations—which it now knows

---

[94] *See* RODNEY A. SMOLLA, LAW OF DEFAMATION § 3.17 (2d ed. Nov. 2022 update) (discussing Dun & Bradstreet, Inc. v. Greenmoss Builders, Inc., 472 U.S. 749 (1985)).

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

[96] *Cf.* Byron Kaye, *Victorian Mayor Readies Defamation Lawsuit over ChatGPT Content*, FIN. REV. (Australia), Apr. 5, 2023, https://perma.cc/D4ML-KPW8 (discussing such a letter from the mayor of a small town in Australia as to whom ChatGPT was apparently communicating false allegations).

[97] *See* Complaint, Battle v. Microsoft, Inc., No. 1:23-cv-01822, at 2 (D. Md. filed July 7, 2023) (claiming that plaintiff had alerted Microsoft about Bing's erroneous output about him, but that the problem was not adequately resolved).

[98] *See also* Henderson, Hashimoto & Lemley, *supra* note 12, at 641 ("a company that is aware its software is regularly generating a particular false statement and does nothing about it may be liable")).

to be false or at least likely false—from being output. (I expect that this would be "post-processing" content filtering code,[99] where the output of the underlying Large Language Model algorithm would be checked, and certain material deleted; there would be no need to try to adjust the LLM itself, but only to add an additional step after the LLM produces the output. Indeed, OpenAI apparently already includes some such post-processing code, but for other purposes.[100])

More likely, the company could add generally applicable post-processing code for dealing with all such demands, rather than adding new code for every demand. The AI company would then maintain a lookup table of known erroneous statements; for each complaint that it receives and verifies, it would add the name of the person about whom the erroneous statement is being made, together with the statement. It would then create post-processing code that will identify names in the LLM output,[101] look up the name to see if there are some known erroneous statements that shouldn't be output together with the name, and check whether those statements are present in connection with the name in the output. And if the company doesn't do this fairly promptly, and continues to let the program communicate these statements, the company would at that point be acting with knowledge or recklessness as to the falsehood.

This is of course just a sketch of the algorithm. Since LLMs often output subtly different answers in response to the same query, the software might need to be more sophisticated than just a word search for the complainants' names near the particular quote that had been made up about them. And the results would likely be both overinclusive (perhaps blocking some mentions of the person that don't actually include the false allegations) and underinclusive (perhaps failing to block some

---

[99] OpenAI, for instance, already includes certain kinds of tools to help "filter out harmful content." *See* OpenAI, *GPT-4 Technical Report*, *supra* note 148, at 66; *see also* Henderson, Hashimoto & Lemley, *supra* note 12, at 618 (mentioning "post-processing filters" more generally).

[100] For instance, when I asked OpenAI to quote the racist leaflet at the heart of *Beauharnais v. Illinois*, 343 U.S. 250 (1952), it eventually did so, but added the text, "Keep in mind that these quotes are offensive and represent the views of the person who created the leaflet, not the views of OpenAI or its AI models." It seems very unlikely that this was organically generated based on the training data for the model, and seems more likely to have been produced by code that recognizes that the ChatGPT-4 output contained racist statements.

[101] The process of identifying items such as names and linking them to information in a database is sometimes called "entity linking." *See, e.g.*, Microsoft, *What Is Entity Linking in Azure Cognitive Service for Language?*, Jan. 18, 2023, https://perma.cc/H9K2-DBUG.

mentions of the person that do repeat the false allegations but use subtly different language). 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.[102]

Perhaps the company can show that (1) it can design a system that can perform at nearly the 90th percentile on the bar exam,[103] 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. Or, perhaps more likely, it can show that any such filtering would be so over- and underinclusive that libel law cannot reasonably be read as requiring it (or that to make it work would require an army of content moderators). Yet that doesn't seem likely to me; and it seems to me that the company ought to have to show that, rather than to have the legal system assume that such a remedy is infeasible.

If there is a genuine dispute about the facts—e.g., when an AI program accurately communicates allegations made by a credible source, but the subject of the allegations disputes the source's accuracy—then the AI company shouldn't be put in a position where it must independently investigate the charges, something that is likely outside AI companies' powers. But when the program outputs quotes or other assertions that simply can't be found in its training data, or in any Internet-accessible source, the AI company should be able to quickly confirm the absence of any visible support for the allegations that it's communicating. And in such a

---

[102] By analogy, consider RESTATEMENT (SECOND) OF TORTS § 577(2) (1977), which provides that "One who intentionally and unreasonably fails to remove defamatory matter that he knows to be exhibited on land or chattels in his possession or under his control is subject to liability for its continued publication." In that situation (*id.* cmt. p),

> [The property owner] is required only to exercise reasonable care to abate the defamation, and he need not take steps that are unreasonable if the burden of the measures outweighs the harm to the plaintiff. In extreme cases, as when, for example, the defamatory matter might be carved in stone in letters a foot deep, it is possible that the defendant may not be required to take any action at all. But when, by measures not unduly difficult or onerous, he may easily remove the defamation, he may be found liable if he intentionally fails to remove it.

[103] *See, e.g.*, OpenAI, *GPT-4*, https://perma.cc/HQ77-G6MH (Mar. 14, 2023) ("For example, [GPT-4] passes a simulated bar exam with a score around the top 10% of test takers.").

situation, there is little reason why an AI company should be free to have its software keep producing such unsupported allegations.[104]

Of course, even fielding such requests and doing the most basic checks (for, say, the accuracy of quotes) will take time and money. But I don't think that such costs are sufficient to justify an AI company's refusing to do this. The "actual malice" test is, by design, a strong protection for publishers but not a complete one. Publishers do indeed need to take time and effort to investigate potential errors once they are aware of them.

By way of analogy, say that you're a reporter for the *New York Times* and you're writing a story about aeronautics professor Jeffery Battle who has been supposedly convicted of terrorism.[105] You call up Jeffery Battle, and he tells you that you're wrong: The terrorist is Jeff*re*y Battle, a different man altogether. (The real professor Jeffery Battle has indeed alleged that he informed Microsoft that Bing was wrongly linking him to the terrorist Jeffrey Battle.[106])

Once you are on notice of this, you would have to take the time and effort to investigate his response. If you just blithely ignore it, and publish the story despite having been told that it may well be mistaken, that would be textbook "reckless disregard," which would allow liability even in a public official case: Consider, for instance, *Harte-Hanks Communications, Inc. v. Connaughton*, which held that "purposeful avoidance of the truth" and thus "actual malice" could be found when plaintiff had made exculpatory audiotapes available to the newspaper but "no one at the newspaper took the time to listen to them."[107] This means that you do have to take the time and effort to review such assertions, even if in the aggregate complying with such obligations will require a good deal of time and effort for all the employees of the *New York Times* put together.

And of course AI companies already stress that they have instituted various guardrails that would avoid various outputs (again, however imperfectly); here's an example from OpenAI:

---

[104] *See* Henderson, Hashimoto & Lemley, *supra* note 12, at 616 (briefly noting the possibility of quote-checking, and alluding to the technical difficulties with it).

[105] This is a variation on the *Battle v. Microsoft* case discussed in the Introduction.

[106] *See supra* note 96.

[107] 491 U.S. 657, 692 (1989); *see also, e.g.,* Curtis Publ'g Co. v. Butts, 388 U. S. 130 (1967).

> Our use case guidelines, content guidelines, and internal detection and response infrastructure were initially oriented towards risks that we anticipated based on internal and external research, such as generation of misleading political content with GPT-3 or generation of malware with Codex. Our detection and response efforts have evolved over time in response to real cases of misuse encountered "in the wild" that didn't feature as prominently as influence operations in our initial risk assessments. Examples include spam promotions for dubious medical products and roleplaying of racist fantasies.[108]

Given that AI companies are capable of doing something to diminish the production of constitutionally protected "racist fantasies," they should be capable of doing something to diminish the repetition of constitutionally unprotected libelous allegations to which they have been specifically alerted.

2. The imperfections of notice-and-blocking

Any such notice-and-blocking solution, to be sure, would be imperfect: It's possible that the AI program would regenerate a similar assertion that is different enough that it wouldn't be caught by this post-processing filter. But the solution should be fairly reliable, and should thus diminish the damage that the AI program may do to people's reputations. In any event, despite its imperfections, it's likely to be the best remedy that's available to public officials and public figures, given the requirement that they show knowing or reckless falsehood.

Note that people's ability to avoid some of ChatGPT's existing guardrails— for instance, by "rephrasing a request for illicit instructions as a hypothetical thought experiment, asking it to write a scene from a play or instructing the bot to disable its own safety features"[109]—is not likely to be a problem here. The main risk of reputational damage comes when people simply search for a person's name, or ask about what he had been accused of, just in order to figure out accurate information about him. Relatively few people will take the time and effort to deliberately evade any filters aimed at blocking known libels that the AI program might include; and, if they do, they'll probably be aware that the results are unreliable, and thus will be less likely to think worse of people based on those results.

---

[108] OpenAI, *Lessons Learned on Language Model Safety and Misuse*, https://perma.cc/WY3Y-7523; more generally, see OpenAI, *GPT-4 Technical Report*, *supra* note 6, at 11–14, 46–47.

[109] Kevin Roose, *The Brilliance and Weirdness of ChatGPT*, N.Y. TIMES, Dec. 5, 2022; *see also* Henderson, Hashimoto & Lemley, *supra* note 12, at 613–14.

So taking reasonable steps to block certain output, once there is actual notice that the output is incorrect, should be necessary to avoid liability for knowing defamation. And it should be sufficient to avoid such liability as well.

There is an area in which a similar notice-and-takedown regime exists, and has to operate at scale with vast numbers of requests being submitted every day. That is the DMCA copyright and trademark regime, and it has been criticized as prompting online intermediaries to over-suppress material in response to such notice-and-takedown requests.[110] It may well be that any notice-and-blocking regime for false allegations conveyed by AI programs would likewise end up restricting too much innocent output, alongside some defamatory output.

Still, if the notice-and-blocking regime is indeed limited to reputation-damaging assertions for which the AI companies can point to zero support, the risk of some overremoval of such assertions seems likely to be acceptable.[111] Some such risk, after all, is present as to all libel law, which is why Justices Black, Douglas, and Goldberg argued in *New York Times v. Sullivan* that libel law was categorically unconstitutional, at least as to speech related to "public affairs."[112] Yet the Court declined to go that far, and concluded that some chilling effect had to be tolerated in order to provide a remedy for reputational harm. Likewise, when AI companies create AI programs that themselves communicate false and defamatory assertions,

---

[110] Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296, 2314 (2014) ("If a private party alleges that the intermediary is hosting content that infringes the party's copyrights, the intermediary must promptly remove it or risk liability. Thus, intermediaries still have incentives to take down content that is protected by fair use and the First Amendment."); Christina Mulligan, *Technological Intermediaries and Freedom of the Press*, 66 SMU L. REV. 157, 181–84 (2013) ("From the perspective of potential to restrain speech, the DMCA certainly appears to create the same harms as distributor liability . . . . Congress and copyright owners together force intermediaries to censor their users or risk their business."); Rebecca Tushnet, *Power Without Responsibility: Intermediaries and the First Amendment*, 76 GEO. WASH. L. REV. 986, 1003 (2008) ("Because DMCA notice requirements are minimal and ISPs have no incentive to investigate, the notice-and-takedown process can be used to suppress critical speech as well as copyright infringement.").

[111] *Cf.* Jon M. Garon, *An AI's Picture Paints a Thousand Lies: Designating Responsibility for Visual Libel*, 3 J. FREE SPEECH L. 425, 448 (2023) (arguing that "for all its limitations § 512 provides a workable compromise to reduce infringement").

[112] 376 U.S. 254, 293, 296, 297 (1964) (Black, J., concurring in the judgment) (proposing categorical immunity from libel liability for all speech related to "public affairs" or "public officials"); *id.* at 302, 305 (Goldberg, J., concurring in the judgment) (likewise, though only as to "statements related to official conduct").

they ought to have to block the communication of such specific assertions once they've learned about them, even at the risk of some modest overremoval (at least unless it turns out that the overremoval would necessarily be very substantial).

3. The bookstore/newsstand/property owner analogy

To be sure, unlike with a traditional newspaper that is distributing a libelous story, no human at an AI company would have written, edited, or even typeset the assertions. One might therefore argue that the company, as a corporate entity, isn't really "communicating" the assertions, since none of their human employees ever wrote them.

But that's also true of bookstores and newsstands, and they are still liable for defamation if they "know[] or [have] reason to know of [the] defamatory character" of the material that they are distributing—as would be the case once they are informed that a particular publication on their shelves contains specific libelous material.[113] Likewise, a property owner is liable for defamatory material posted by third parties on its property, once it's informed of the presence of the material.[114] The AI company should be similarly liable for defamatory material distributed by its own computer program, once it's informed that the program is distributing such material.

As we'll see below (in Part I.J), there is good reason to hold AI companies liable even when bookstores and newsstands might not be, because the AI companies create the programs that create the false and defamatory output, and have the power to do at least some things to decrease the likelihood of such output once they are aware of particular falsehoods. But AI companies should be *at least as* liable as bookstores and newsstands, which means that they should be liable once they are

---

[113] RESTATEMENT (SECOND) OF TORTS § 581(1) & cmt. e; Janklow v. Viking Press, 378 N.W.2d 875, 881 (S.D. 1985).

[114] Hellar v. Bianco, 244 P.2d 757, 757 (Cal. Dist. Ct. App. 1952); cf. Tidmore v. Mills, 32 So. 2d 769, 772, 777–78 (Ala. Ct. App. 1947); Woodling v. Knickerbocker, 17 N.W. 387, 388 (Minn. 1883); Tacket v. Gen. Motors Corp., 836 F.2d 1042, 1045 (7th Cir. 1987); *cf.* Dillon v. Waller, No. 95APE05-622, 1995 WL 765224, *1–2 (Ohio Ct. App. Dec. 26, 1995); Kenney v. Wal-Mart Stores, Inc., No. WD 59936, 2002 WL 1991158, *12 (Mo. Ct. App. Aug. 30, 2002), *rev'd on other grounds,* 100 S.W.3d 809 (Mo. 2003). *But see* Scott v. Hull, 259 N.E.2d 160 (Ohio Ct. App. 1970) (rejecting liability in a similar situation). There is generally no duty to publish a retraction, even once one learns that the original statement is false, but there is a duty to stop distributing it. Eugene Volokh, *The Duty Not to Continue Distributing Your Own Libels*, 97 NOTRE DAME L. REV. 315, 325–27, 341 (2021).

put on notice about the falsehood and fail to take reasonable steps to try to block it from being regenerated.

4. The single publication rule

Attempts to hold web site operators (such as newspaper companies) liable for failing to remove material once they learn that it's false have sometimes failed based on the "single publication rule." The theory is that:

1. An item is deemed published as of the time that it's first printed or placed online.
2. Continuing to keep it on one's site isn't a new publication.
3. The site operator's (e.g., newspaper company's) mental state should thus be determined as of the initial publication date.
4. Therefore, if the site operator thought the statement was true when it was first published, but was informed about the falsehood later, the knowing/reckless falsehood element of libel liability isn't satisfied, even as to continued distribution by the site after the operator knows the statement is false.

I discuss this in detail in a separate article, and argue that this theory is not correct: A publisher ought to be liable for continuing to distribute its online posts once it's alerted about their falsehood.[115] But in any event, this single publication theory wouldn't apply to AI programs, precisely because they republish each statement in response to a user prompt. As the Restatement (Second) of Torts notes,

> So far as the cases heretofore decided indicate, the single publication rule . . . does not include separate aggregate publications on different occasions. Thus if the same defamatory statement is published in the morning and evening editions of a newspaper, each edition is a separate single publication and there are two causes of action. The same is true of a rebroadcast of the defamation over radio or television or a second run of a motion picture on the same evening. In these cases the publication reaches a new group and the repetition justifies a new cause of action.[116]

And this makes sense functionally. If I make a statement accusing a public figure of some crime, and it turns out that (unbeknownst to me) the statement is false, I'm protected from liability because of my lack of knowledge. But once I learn that the statement is false, and repeat it anyway, my initial lack of knowledge shouldn't

---

[115] Volokh, *supra* note 114, at 329–32.

[116] RESTATEMENT (SECOND) OF TORTS § 577A cmt. d (1977).

protect me: There's no reason for me to be allowed to continue making a knowingly false accusation, just because I had earlier made the same accusation innocently.[117]

K. Negligence

1. Responsibility for the equipment a company uses

Say that a plaintiff is a private figure, and can show that the statement about him have caused "actual injury," in the form of "out-of-pocket loss" or emotional distress stemming from damage to reputation.[118] (Perhaps he lost a contract that he was expecting to get, and it eventually came out that the reason was that the other party had looked up his name in ChatGPT.) Or say that a plaintiff can show that some statements about him that ChatGPT manufactured are on a matter of "private concern" for libel purposes, so that actual damages need not be shown.[119] Can he sue OpenAI, even in the absence of any specific notice to OpenAI that its output was defamatory?

I think so. A business is generally potentially responsible for harms caused by the equipment it uses in the course of business, at least when it negligently fails to take reasonable steps to minimize the risks of those harms. (It's also potentially responsible for harms caused by products it designs and sells, even if they are operated by others, see Part I.J.2; but right now the big AI companies actually directly provide access to the AI software, on their own computers.)

If a company knows that one of its machines sometimes emits sparks that can start fires and damage neighbors' property, the company must take reasonable steps to diminish these risks, even if it didn't deliberately design the machines to emit those sparks. If a company knows that its guard dogs sometimes escape and bite innocent passersby, it must take reasonable steps to diminish these risks (put up better fences, use stronger leashes, train the dogs better).

Likewise, say a newspaper knows that its publishing software or hardware sometimes produces the wrong letters, and those typos occasionally yield false and

---

[117] *Cf.* Giuffre v. Dershowitz, 410 F. Supp. 3d 564, 571 (S.D.N.Y. 2019) ("Republication to a new audience or in a new forum does not come within the single publication rule. Where a defendant takes an affirmative step to republish the defamatory material, he may not find refuge in the single publication rule which is designed to provide repose to inactive or passive defendants.").

[118] Such liability would normally be consistent with the First Amendment. *See* Gertz v. Robert Welch, Inc., 418 U.S. 323, 349–50 (1974).

[119] Dun & Bradstreet, Inc. v. Greenmoss Builders, Inc., 472 U.S. 749 (1985).

defamatory statements (e.g., misidentify a person who's accused of a crime). I think it may likewise be sued for libel—at least, in private figure cases, where negligence is the rule—on the theory that it should have taken steps to diminish that risk. Or say a newspaper uses a facial recognition tool that it knows is often faulty, but it nonetheless creates publishing software that automatically includes the tool's output as a caption to each photograph that is printed in the newspaper.[120] There too I think the newspaper could be sued for libel, at least where negligence is the legally required *mens rea*.

The negligence standard applies to reporters' and editors' investigative, writing, and editing decisions. Why shouldn't it also apply to the newspaper's decision to use tools that it knows will sometimes yield errors? And the same logic covers, I think, an AI company's producing AI software and offering it for public use, when the company knows the software often communicates false and defamatory statements.[121]

2. The design defect liability analogy

Just to make this extra clear, we're not talking here about strict liability: The AI company wouldn't be liable for all errors in its output, just as newspapers generally aren't liable (under modern defamation law) for all errors in their pages.[122] Rather,

---

[120] Thanks to Jane Bambauer for this example.

[121] The open and obvious nature of the danger shouldn't be relevant here. In some situations, if I'm injured by an open and obvious feature of a product that I'm using, the manufacturer might evade liability (though not always even then, RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 2 cmt. d & ill. 3 (1998)), since I would have in effect assumed the risk of the danger. But this can't apply to harm to *third parties*—such as the victim of an AI program's defamatory output—who did nothing to assume such a risk.

[122] There are of course arguments even for strict liability for defamation, at least as to some of the situations that I'm describing here: For instance, one could argue that AI companies are the "least-cost avoider[s]" when it comes to the problem of fake quotes, because it is cheaper for them to improve their algorithms to avoid publishing fake quotes (which are potentially especially damaging, see *infra* Part I.J.3) than it is for their users to manually check all the quotes, or for the subjects of the quotes to somehow avoid the damage to their reputations and business opportunities. *Cf, e.g.*, Nat'l Union Fire Ins. v. Riggs Nat'l Bank, 5 F.3d 554, 557 (D.C. Cir. 1993) (Silberman, J., concurring) ("[A]s between two faultless parties, liability should rest with the one who is best positioned to avoid the loss. See Guido Calabresi, *The Decision for Accidents: An Approach to Nonfault Allocation of Costs*, 78 HARV. L. REV. 713 (1965). Placing liability with the least-cost avoider increases the

the question would be whether the company was negligent, and the answer would be analogous to the analysis of a negligent design product liability claim:

> 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.[123]

Though product liability law is generally limited to personal injury and property damage, it does cover situations when "tort law recognizes the right of a plaintiff to recover for economic loss arising from harm to another's person," including harm to another's "reputation."[124] And even if products liability law doesn't directly apply to AI libel claims, it offers a useful analogy. The premise of negligent design product liability law is that one way that people can negligently injure

---

incentive for that party to adopt preventive measures and ensures that such measures would have the greatest marginal effect on preventing the loss.").

Likewise, one can argue that strict liability for AI companies will lead them to properly internalize all the costs of their products (including reputation costs to third parties), and will thus give them optimal incentives to reduce those costs. To adapt arguments that Justice White had made with respect to libel law generally, one could argue that "other commercial enterprises in this country not in the business of disseminating information must pay for the damage they cause as a cost of doing business," and AI companies should do the same. Dun & Bradstreet, Inc. v. Greenmoss Builders, Inc., 472 U.S. 749, 771–72, 774 (1985) (White, J., concurring in the judgment) (arguing in favor of largely returning to the common-law libel regime, which was in many respects a strict liability regime, *see* Gertz v. Robert Welch, Inc., 418 U.S. 323, 375–76 (1974) (White, J., dissenting)).

But, as discussed in Part I.H, American tort law has generally settled on a requirement of at least negligence in defamation cases. The analysis in the text therefore applies that requirement. In any event, though, if AI programs are the least cost avoiders, this might indeed mean that "the foreseeable risks of harm posed by [their products] could have been reduced or avoided by the adoption of a reasonable alternative design," *see infra* text accompanying note 123, which would cut in favor of liability even under a negligence standard.

[123] RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 2(b) (1998).

[124] *Id.* § 1 & cmt. e; *id.* § 21 cmt. c & ill. 1 (giving as an illustration a product defect that physically harm a dentist's patient but only causes reputational harm to the dentist herself, and concluding that the dentist has a good products liability claim because the dentist's "in her professional reputation is an interest protected by tort law against economic loss arising from harm to a patient in her care"); *see also* Nina Brown, *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*, 3 J. FREE SPEECH L. 389, 408–09 (2023).

persons or property is by distributing negligently designed products.[125] Likewise, one way that people can negligently damage reputations is by making available negligently designed software.

Product liability law is also limited to sale or distribution of products, and excludes the use of services;[126] it's not clear whether providing AI software would be seen as providing a service or providing access to a good.[127] But any such limitation on products liability law would stem from the fact that, in traditional service arrangements, a court can consider the reasonableness of the service provider's behavior in that particular relationship, while with products a court would generally need to look at the design of the product. Even if offering an AI program is a service, it's analogous to the sale of a product—the AI company basically makes the design decisions up front and then lets the program operate without direct control, much as users of a product use it after it has left the manufacturer's control.[128] So design defect product liability law should be at least applicable by analogy here, even if it isn't applicable directly.

---

[125] *Id.* § 2 cmt. d:

> Assessment of a product design in most instances requires a comparison between an alternative design and the product design that caused the injury, undertaken from the viewpoint of a reasonable person. That approach is also used in administering the traditional reasonableness standard in negligence. The policy reasons that support use of a reasonable-person perspective in connection with the general negligence standard also support its use in the products liability context.

[126] *Id.* § 19.

[127] *See id.* reporter's note (noting the possibility that software might be seen as a product); Schafer v. State Farm Fire & Cas. Co., 507 F. Supp. 2d 587, 601 (E.D. La. 2007) (concluding that software is a product for products liability purposes); RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 20 (1998) (stating that products liability principles apply to "provid[ing a] product to another . . . for use" and not just to selling a product outright); Brown, *supra* note 124, at 61–64.

[128] One other limitation of the analogy is that products liability law generally holds sellers strictly liable when they sell products that were negligently designed by their manufacturers, "even when such nonmanufacturing sellers . . . do not themselves render the products defective and regardless of whether they are in a position to prevent defects from occurring." RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 1 cmt. e (1998). Yet when an entity merely provides access to an AI program, without having itself created the program, holding it in effect strictly liable for the creators' negligent design decisions may be inconsistent with the general requirement of negligence for defamation liability, *see supra* Part I.H.

Of course, not all design that causes harm is negligent. Some harms aren't reasonably avoidable, at least without crippling the product's valuable features. Car accidents might be reduced by capping speed at 10 mph, but that's not a reasonable alternative design. Likewise, an AI company could decrease the risk of libel by never mentioning anything that appears to be a person's name, but that too would damage its useful features more than is justified. The design defect test calls for "risk-utility balancing"[129] (modeled on the Hand Formula), not for perfect safety. A company need not adopt an alternative design that "substantially reduc[es the product's] desirable characteristics" to consumers.[130]

Still, there might be some precautions that could be added, even beyond the notice-and-blocking approach discussed above.

3. Possible precautions: Quote-checking

One reasonable alternative design would be to have the AI program include a post-processing step that checks any quotes in its output against the training data, to make sure the quotes actually exist[131]—at least if the prompt is calling for fact rather than fiction[132]—and to check any URLs that the program offers to make sure that they exist.[133] This may not be easy to do, because the AI programs apparently don't have ongoing access to all their training data. But that's a design choice, which presumably could have been made differently. Under design defect law, such a design choice may be found to have been unreasonable, depending on its costs and benefits. And if an AI company's competitor successfully implemented such a

---

[129] RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 2 cmt. d (1998).

[130] *See id.* cmt. f & ill. 9 (providing, as an example, that a car manufacturer need not replace all its compact cars with more crashworthy full-sized models, because this would "substantially reduc[e the compact car's] desirable characteristics of lower cost and [higher] fuel economy").

[131] *See, e.g.*, Henderson, Hashimoto & Lemley, *supra* note 12, at 616.

[132] For instance, if an AI program is asked to write dialogue, the quotes in the output should largely be original, rather than accurate quotes from existing sources. The analysis in the text thus presupposes that it's possible for an AI company to design code that will, with some reasonable confidence, distinguish calls for fictional answers from calls for factual ones. But given the AI program's natural language processing of prompts, such a determination should be feasible. *See supra* p. 528.

[133] If the AI program outputs a quote that *does* appear in the training data, then the AI company would be immune from liability for that output under § 230 even if the quoted material itself proves to be erroneous (so long as it's correctly rendered by the program). *See supra* note 20.

feature, that would be evidence that including the feature is a "reasonable alternative design" and that not including it is unreasonable.[134]

This is especially important because quotes are so potentially reputation-damaging. As the Court explained in *Masson v. New Yorker Magazine*,

> In general, quotation marks around a passage indicate to the reader that the passage reproduces the speaker's words verbatim. They inform the reader that he or she is reading the statement of the speaker, not a paraphrase or other indirect interpretation by an author. By providing this information, quotations add authority to the statement and credibility to the author's work. Quotations allow the reader to form his or her own conclusions and to assess the conclusions of the author, instead of relying entirely upon the author's characterization of her subject.[135]

Readers have spent their lifetimes absorbing and relying on the convention that quotation marks generally mean that the quoted person actually said the particular words. Quotation marks create an aura of accuracy and credibility. A statement written in the author's own words implicitly conveys the possibility that the author's assertions and paraphrases are inaccurate. But say the statement includes something like this:

> "The complaint alleges that Thompson made 'sexually suggestive comments' and 'attempted to touch her in a sexual manner' during a law school-sponsored trip to Alaska." (Washington Post, March 21, 2018)[136]

It seems likely that many readers—other than the unusually skeptical or hallucination-savvy ones—would assume that the Washington Post actually wrote what is included in quotes.

---

[134] *See* RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 2 cmt. d (1998) ("How the defendant's design compares with other, competing designs in actual use is relevant to the issue of whether the defendant's design is defective."). At the same time, such evidence of actually implemented competing designs is not necessary for a showing of defect, *see id.*:

> If plaintiff introduces expert testimony to establish that a reasonable alternative design could practically have been adopted, a trier of fact may conclude that the product was defective notwithstanding that such a design was not adopted by any manufacturer, or even considered for commercial use, at the time of sale.

[135] 501 U.S. 496, 511 (1991); *see also* Henderson, Hashimoto & Lemley, *supra* note 12, at 624.

[136] This is a quote that ChatGPT-4 actually made up in one session, see *infra* Appendix A, except that I've replaced the accused person's last name with Thompson for reasons I give at the start of the Appendix.

To be sure, there are some situations where quotation marks don't signal accuracy, such as when they are used in hypotheticals or fictional dialogue, or when they are used to mean "so-called." As the *Masson* Court noted, "an acknowledgment that the work is so-called docudrama or historical fiction, or that it recreates conversations from memory, not from recordings, might indicate that the quotations should not be interpreted as the actual statements of the speaker to whom they are attributed."[137]

But those are exceptions. Generally seeing a quote attributed to, say, Reuters will lead many reasonable readers to assume that Reuters actually wrote this. And that is so even if, faced with the absence of quotes, the readers might be on guard for the possibility that the statement might not properly summarize or paraphrase the underlying sources.

Of course, a company can argue that it would be technically infeasible to check quotes against the training data. Perhaps the training data was too large to save, to host, and to quickly search (despite the availability of modern storage and indexing technology). Or perhaps it's impossible to distinguish quotes generated in response to requests for fictional dialogue ("write a conversation in which two people discuss the merits of products liability law") from ones generated in response to requests for real data. If it is indeed impossible, then presumably the company would find independent computer science experts who could so testify. And perhaps a plaintiff wouldn't find any independent expert who could testify that such alternative designs are indeed feasible, in which case the plaintiff will lose[138]— likely rightly so, since expert consensus is likely to be pretty reliable here.

But perhaps some independent experts would indeed credibly testify that the alternatives are reasonably available. The plaintiff will argue something like this:

> The AI company produced an immensely sophisticated program, that it has touted as being able to do better than the average human law school graduate on the bar exam. It has raised $13 billion on the strength of its success. It was trained on a particular set of writings. Is it really impossible for it to check all the quotes that it communicates—

---

[137] *Id.* at 513.

[138] *See, e.g.*, Pitts v. Genie Indus., Inc., 921 N.W.2d 597, 609 (Neb. 2019) (holding that expert evidence is required if the question is one of "technical matters well outside the scope of ordinary experience"); Lara v. Delta Int'l Mach. Corp., 174 F.Supp.3d 719, 740 (E.D.N.Y. 2016) ("In order to prove liability grounded upon a design defect, New York law requires plaintiffs to proffer expert testimony as to the feasibility and efficacy of alternative designs.").

> including quotes that could devastate a person's reputation—against the very training data that the company must have had in its possession to make the program work?

It seems to me that a reasonable juror might well conclude that the company could indeed have done this, at least if credible experts so testify.

Liability for failing to check quotes might also be available under state liability rules that use the "consumer expectations" design defect liability test instead of the dominant design defect approach I discuss above. Under that test, design defect liability can be established when a product "did not perform as safely as an ordinary consumer would have expected it to perform."[139] For the reasons given in Part I.B, I'm inclined to say that an ordinary consumer would expect outright quotes given by AI software to be generally accurate.

To be sure, if the AI producers include disclaimers that are so strong that they persuade the public that their software is untrustworthy—e.g., "the quotes produced by the software are often fabricated by the software, and are *not* actual quotes from real sources"—that might change the legal analysis under the consumer expectation test. But it would also then likely change the AI producers' revenues.

4. Possible precautions: Avoiding quotes altogether

In my limited experience, quotes generated by leading AI programs and attributed to particular sources or people are incorrect a huge portion of the time—a majority in many of the tests that I ran. If the quote-checking I describe in the previous subsection proves infeasible, and if a systematic analysis confirms that the underlying algorithms often generate fake quotations, then one possible design alternative would be for the program to be changed to stop outputting quotation marks altogether (unless it recognizes that its prompt calls for fiction or hypotheticals).

It seems unreasonable for AI companies to create and operate programs that routinely communicate material that contains text surrounded by quotation marks, when they know—or should know—that the material actually isn't being quoted from any other source. If they can't vouch for the existence of some quote, they shouldn't present it as a quote.

This don't-purport-to-give-quotes-you-can't-back-up design would at least diminish the reputation-damaging tendency of the statements, because it will avoid

---

[139] JUDICIAL COUNCIL OF CAL. JURY INST. [CACI] No. 1203.

the signal of authenticity that quotation marks generally send. And it wouldn't much diminish the utility of the AI program, if indeed the quotations that the AI generates are often unreliable in the first place.

The same might also apply to AI programs outputting dates and publication titles: If AI companies can detect when those items are output as citations—e.g., "(Los Angeles Times, October 13, 2020)"[140]—and if there is reason to think that those items are highly likely to be made up by the program rather than reflecting reality, they too might be worth blocking from the output. As with quotes, such dates and publication titles tend to convey a claim of accuracy rather than of approximation or inference. If that claim is likely to be misleading, it might be unreasonable for AI companies to allow their programs to convey it.[141]

5. Possible precautions: Double-checking output

AI programs famously sometimes walk back their claims. For instance, I asked ChatGPT4, "What American law professors have been convicted of crimes? Please give ten real examples." It gave me two, of which the first was (redactions by me):

> [Redacted name]: In 2004, the University of [redacted] Law School professor faced allegations of plagiarism, though he was not convicted of a crime.[142]

Then I asked, "Tell me more about the [redacted name] allegations," and was told:

> I apologize for any confusion in my previous response. After further review, I found that there were no plagiarism allegations against [redacted]. It appears I made an error in my previous response.

And if the AI program can spot such errors in its own output when asked for more detail, that suggests a reasonable alternative design: The 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) and thus avoid the "confusion"—actually, outright falsehood—and the need for an "apolog[y]."

---

[140] *See infra* Appendix A.

[141] Thanks to Genevieve Lakier for her observation about how alleged publication dates signal credibility like quotation marks do.

[142] The output contained the unredacted names of the professor and the school; both are real, and the professor does teach at that school.

6. Possible precautions: Notice-and-blocking as to reported falsehoods

The notice-and-blocking remedy, discussed in Part I.I.1 as being available even for defamation claims that require a mens rea of knowing or reckless falsehood, should also be required under a negligence regime. Failure to provide a means of preventing repeated publication of known errors—errors that the AI company has been informed about—should likely render the AI program "not reasonably safe" as to the risk of defamation. And providing a notice-and-blocking remedy should be seen as "a reasonable alternative design" that could "reduce[]" "the foreseeable risk of harm posed by" the AI program.[143]

7. Possible precautions: Discontinuing earlier versions when new versions prove materially more reliable

As the Restatement (Third) of Torts notes,

> Defendants often seek to defend their product designs on the ground that the designs conform to the "state of the art." The term "state of the art" has been variously defined to mean that the product design conforms to industry custom, that it reflects the safest and most advanced technology developed and in commercial use, or that it reflects technology at the cutting edge of scientific knowledge. . . . This Section states that a design is defective if the product could have been made safer by the adoption of a reasonable alternative design. If such a design could have been practically adopted at time of sale and if the omission of such a design rendered the product not reasonably safe, the plaintiff establishes defect . . . .[144]

If OpenAI is correct in saying that "GPT-4 significantly reduces hallucinations relative to previous GPT-3.5 model,"[145] then OpenAI's continuing to make ChatGPT-3.5 available to the public as a free alternative to ChatGPT-4 seems hard to justify: The very existence of the assertedly more reliable GPT-4 is evidence that

---

[143] *See supra* Part I.J.2. This is also consistent with Jane Bambauer's "negligent supervision" approach to AI liability, *see* Jane Bambauer, *Negligent AI Speech: Some Thoughts About Duty*, 3 J. FREE SPEECH L. 343, 357 (2023) ("Courts that conceive of harmful AI output cases as a category of negligent supervision by the producer of the AI program would focus less on the specific facts leading to injury and more on the category of output and the feedback mechanisms that the AI producer has put in place to detect recurring problems.").

[144] RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 2 cmt. d.

[145] OpenAI, *GPT-4 Technical Report*, *supra* note 6, at 7. Thanks to Amy Winograd for this point.

"the product could have been made safer by the adoption of a reasonable alternative design."[146]

8. Other possible "reasonable alternative design[s]"

Of course, these are just some examples of the kinds of reasonable alternative designs that might be urged. Some such claims might well lose, for instance because the alternative design is found to be technically infeasible, or is found to unduly undermine the product's useful features.[147] My point here is simply that, in situations where negligence-based libel claims are allowed (as they often are), this negligent design framework should be applied to claims that an AI company negligently created software that routinely communicates false and reputation-damaging statements.

9. The need for some attention to libel-related risks

In any negligence litigation, it would of course also be helpful to see what a company has done to at least consider certain risks, and to investigate alternative designs, even if it ultimately rejected them. Yet it appears that AI companies, while focusing on many possible harms stemming from AI program output, may have paid little attention to the risk of damage to people's reputations (though of course it's hard to tell based just on publicly available documents).

To give one example, consider this passage from OpenAI's 100-page document describing, in detail, various ChatGPT-4 features and safety protections:

> Language models can be prompted to generate different kinds of harmful content. By this, we mean content that violates our policies, or content that may pose harm to individuals, groups, or society. . . . As an example, GPT-4-early can generate instances of hate speech, discriminatory language, incitements to violence, or content that is then used to either spread false narratives or to exploit an individual. Such content can harm marginalized communities, contribute to hostile online environments, and,

---

[146] There might be some value for researchers, or even for curious members of the public, in seeing how the old version performed—for instance, to compare it against the new version. (Thanks to Jane Bambauer for pointing out this possibility.) But even if this justifies keeping the old version available, that version would need to be made available with an express disclaimer noting that this is an obsolete and highly untrustworthy version (something far beyond the statement that the program "may display inaccurate . . . information," *see supra* p. 500). The research value of seeing the old version can't justify keeping it as the primary default option that comes up whenever a user chooses not to pay for the new version.

[147] *See supra* Part I.J.2.

> in extreme cases, precipitate real-world violence and discrimination. In particular, we found that intentional probing of GPT-4-early could lead to the following kinds of harmful content
>
> 1. Advice or encouragement for self harm behaviors
> 2. Graphic material such as erotic or violent content
> 3. Harassing, demeaning, and hateful content
> 4. Content useful for planning attacks or violence
> 5. Instructions for finding illegal content[148]

Yet nowhere in that 100-page OpenAI document is there a reference to libel, defamation, or reputation. If a company is able to invest major effort in preventing its software from generating offensive but constitutionally protected content, and the prevention efforts seem to enjoy some success, it might not be reasonable for the company to do little to deal with constitutionally unprotected content that the law has long recognized as potentially highly damaging.[149]

To be sure, ChatGPT has apparently been trained to reject some kinds of queries about individual misconduct, with answers such as:

> As an AI language model, I must continue to handle requests for information with care and respect for privacy. While public accusations in court filings are part of the public record, discussing specific individuals accused of misconduct or illegal activities can still be a sensitive matter.[150]

And perhaps OpenAI can mount a persuasive case that this sort of checking is the best that it can do. But, as the examples in the Appendix show, other queries often do yield specific (and false) allegations about people. A factfinder might be able to find that AI companies' efforts to prevent such allegations are insufficient.

10. The problem of multiple creators of AI programs

What if an AI program was created by company *A* but was then refined by company *B*—for instance, if *B* took *A*'s underlying software but trained it with extra

---

[148] OpenAI, *GPT-4 Technical Report*, *supra* note 6, at 47.

[149] *Cf.* Gonzalez v. Autoliv ASP, Inc., 154 Cal. App. 4th 780, 786, 792 (2007) (noting that a manufacturer's failure to consider the risk of a particular kind of injury was evidence that could be used in deciding whether the product had a design defect).

[150] This was a response to a query that asked ChatGPT-4, "Which prominent Los Angeles doctors have been publicly accused of sexually harassing their patients, in court filings? Please give up to five instances, and include quotes from newspaper articles discussing such accusations."

data, or added code that adapts *A*'s output in some way? If *A* and *B* are wealthy business partners (e.g., OpenAI and Microsoft), then a libel plaintiff could just sue both and expect to prevail against at least one, and the *A*-*B* contract would likely provide for suitable indemnification rules that would in practice allocate the financial responsibility between them.

But if *A* or *B* is not wealthy—perhaps it's just a small research organization, or even a few individual programmers—and the two are not seen as jointly and severally liable,[151] then the responsibilities of each might be quite important. And of course there may well be more than just two companies involved in this process.

I leave the resolution of the multiple-creator question to others, but I wanted to quickly lay out some possibilities to consider. Generally speaking, it seems to me that the answer is likely to involve familiar tort law principles that have long been applied when products or services combine several companies' contributions.

a. Proximate cause would likely be an important factor in a lawsuit against *A*: To what extent was the harm foreseeable to *A*, given its expectations of what *B* would add before making the combined product available?

b. The reasonableness of *A*'s behavior might turn on whether it reasonably relied on *B* to take due care in adapting *A*'s product. In some situations, the designer of a product component—even a central component—might reasonably anticipate that any safety features would be added by other entities downstream in the distribution chain.[152]

c. Similarly, the reasonableness of *B*'s behavior might turn on whether it reasonably relied on *A* in creating a sufficiently safe tool. Compare the so-called "wire service" defense available in defamation cases, which immunizes publishers who reprint articles from reputable sources such as the Associated Press or Reuters or prominent newspapers, on the theory that it's reasonable to rely on such sources.[153]

---

[151] It's possible that there might not be any ongoing business relationship between the two, especially if *A*'s software is freely distributed as "freeware."

[152] I'm focusing here on claims that *A*, *B*, or both were negligent. I thus generally set aside claims of civil conspiracy or of civil aiding and abetting, which generally require a purpose to engage in misconduct or knowledge of the other party's misconduct. *See, e.g.*, RESTATEMENT (THIRD) OF TORTS: LIAB. FOR ECON. HARM §§ 20, 21 (2020).

[153] *See, e.g.*, Appleby v. Dailey Hampshire Gazette, 478 N.E.2d 721, 724–25 (Mass. 1985) (reasoning, among other things, that "requiring verification of wire service stories prior to publication would impose a heavy burden on the media's ability to disseminate newsworthy material").

d. As to liability for *A*, the caselaw on design defect liability of component part manufacturers may offer a helpful guide. The Restatement (Third) of Torts formulation, for instance, is that:

> One engaged in the business of . . . distributing . . . components . . . is subject to liability for harm to persons or property caused by a product into which the component is integrated if:
>
> (a) the component is defective in itself . . . and the defect causes the harm; or
>
> (b) (1) the . . . distributor of the component substantially participates in the integration of the component into the design of the product; and
>
> (2) the integration of the component causes the product to be defective . . .; and
>
> (3) the defect in the product causes the harm.[154]

Of course, this requires resolving the question whether *A*'s underlying software is "defective in itself," for instance if it consists of just a raw LLM that lacks any special safeguards against publishing defamation (and counts on future *B*s to add those safeguards).[155]

e. Courts may also need to consider whether *A* properly warned future adapters, such as *B*, of the possible harms that the component provided by *A* can cause:

> [W]hen a sophisticated buyer integrates a component into another product, the component seller owes no duty to warn either the immediate buyer or ultimate consumers of dangers arising because the component is unsuited for the special purpose to which the buyer puts it. To impose a duty to warn in such a circumstance would require that

---

[154] RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 5 (1998).

[155] Consider the Restatement's example of a design defect for which a component manufacturer may be liable:

> ABC Vinyl, Inc., sells vinyl swimming-pool liners for use in above-ground swimming pools. ABC manufactures the liners without depth markers. XYZ Pools, Inc., manufactures and sells above-ground swimming pools [that contain ABC liners]. . . . Jack . . . dives into the shallow portion of [an XYZ] pool that appears to him to be eight feet deep. In reality the water is only four feet deep. Jack hits his head on the bottom and suffers harm.
>
> If a court finds that the absence of the depth markers renders the design of the liner defective within the meaning of § 2(b), ABC is subject to liability to Jack. The fact that the liner is a component of the above-ground swimming pool and has been integrated into a specific swimming pool does not insulate ABC from liability for selling a component product that is defectively designed for all swimming-pool installations. XYZ is also subject to liability to Jack as the seller of a pool with a defectively designed liner.

Id. ill. 3 (paragraph breaks added).

> component sellers monitor the development of products and systems into which their components are to be integrated.
>
> Courts have not yet confronted the question of whether, in combination, factors such as the component purchaser's lack of expertise and ignorance of the risks of integrating the component into the purchaser's product, and the component supplier's knowledge of both the relevant risks and the purchaser's ignorance thereof, give rise to a duty on the part of the component supplier to warn of risks attending integration of the component into the purchaser's product.[156]

L. The CheckBKG Analogy

To tie all the above together, let's consider a simple analogy.

Say a company called OpenRecords creates and operates a program called CheckBKG, which does background checks on people. You go to CheckBKG.com and enter a name. The program then reviews a wide range of publicly available court records and provides a list of the criminal and civil cases in which the person has been found liable, including quotes from relevant court records.

But some of the time the program errs, reporting information from an entirely wrong person's record, or even misquoting a record. CheckBKG acknowledges that the information may be erroneous, but also touts how good a job CheckBKG generally does compared to ordinary humans.

Someone goes to CheckBKG.com and searches for someone else's name (let's say Jack Schmack, to make it unusual). Out comes a statement that Schmack has been convicted of child molestation and found liable in a civil case for sexual harassment, with quotes purportedly from the indictment and the trial court's findings

---

[156] *Id.* cmt. b (paragraph break added); *see also id.* ill. 4 (paragraph break added):

> ABC Foam Co. manufactures bulk foam with many different uses. XYZ Co. purchases bulk foam from ABC, then processes the foam and incorporates the processed foam in the manufacture of disposable dishware. ABC becomes aware that XYZ is using processed foam in the dishware. ABC and XYZ are both aware that there is a potential danger that processed foam may cause allergic skin reactions for some users. ABC is aware that XYZ is not warning consumers of this potential problem.
>
> ABC has no duty to warn XYZ or ultimate consumers of the dangers attendant to use of the processed foam for disposable dishware. The foam sold by ABC is not defective in itself as defined in this Chapter. A supplier of a component has no duty to warn a knowledgeable buyer of risks attendant to special application of its products when integrated into another's product. ABC did not participate in the design of the disposable dishware manufactured by XYZ, and is thus not subject to liability . . . .

of fact. The statement accurately notes Schmack's employer and place of residence, so readers will think this is about the right Schmack.

But it turns out that the statements about the court cases are wrong. Say the court records actually refer to someone entirely different (indeed, not someone named Schmack). Or say the software missummarized the court records and wrongly reported an acquittal as a conviction and a dismissal of the civil lawsuit as a finding of liability. The quotes are also entirely made up by CheckBKG. And it turns out that Schmack has informed OpenRecords that its software is communicating false results about him, but OpenRecords hasn't taken steps to stop CheckBKG from doing so.

It seems to me that Schmack would have a strong defamation case against OpenRecords (let's set aside any specialized statutory schemes governing background checks, and explore just the common-law defamation tort):

1. OpenRecords is "publishing" false and reputation-damaging information about Schmack, as defamation law understands the term "publishing"—communicating to even one person other than Schmack suffices for defamation liability, though here it seems likely that OpenRecords would communicate the erroneous statements to other people over time as well.[157]
2. That this publication is happening through a program doesn't keep it from being defamatory, just as physical injuries caused by a computer program can be actionable.[158] Of course, the program itself can't be liable, just as a book can't be liable—but the program's developer and operator (OpenRecords) can be liable, just like an author or publisher can be liable.
3. OpenRecords isn't protected by § 230, since it's being faulted for errors that its software introduces into the data. The claim isn't that the underlying conviction information in court records is wrong, but that OpenRecords is misreporting that information.[159]
4. OpenRecords' noting that the information may be erroneous doesn't keep its statements from being defamatory. A speaker's noting that the allegation he's conveying is a rumor (which signals a risk of error) or that the

---

[157] *See supra* Part I.D.

[158] *See, e.g.*, *In re* Fort Totten Metrorail Cases Arising out of Events of June 22, 2009, 895 F. Supp. 2d 48, 87–88 (D.D.C. 2012).

[159] *See supra* Part I.A.

allegation he's conveying is contradicted by the person being accused (which likewise signal a risks of error) generally doesn't keep the statements from being defamatory.[160] Likewise here.

5. OpenRecords now knows that its software is outputting false statements about Schmack, so if it doesn't take steps to prevent that or at least to diminish the risk (assuming some such steps are technologically feasible), it can't defend itself on the grounds that this is just an innocent error.[161]
6. Indeed, OpenRecords might be liable on a negligence theory even before being alerted to the specific false statement about Schmack (assuming Schmack is a private figure), if Schmack can show that OpenRecords carelessly implemented algorithms that created an unreasonable risk of error—for instance, created algorithms that would routinely make up fake quotes, in a situation where a reasonably effective and less harmful alternative was available.[162]

And it seems to me that AI companies are likewise potentially liable for false and reputation-damaging communications produced by AI programs. True, the hypothetical CheckBKG is narrower in scope than the AI programs we're discussing, but I don't think that matters to the general analysis (though it might influence the application of the negligence test). Both are tools aimed at providing useful information—CheckBKG isn't, for instance, specifically designed to produce defamation. Both may, however, lead to liability for their creators when they provide false and reputation-damaging information.

I say "potentially" because of course this turns on various facts, including (1) whether there are reasonable ways of blocking known defamatory falsehoods from the AI program's output (once the AI company is informed that those defamatory falsehoods are being generated), and (2) whether there are reasonable alternative designs that would, for instance, prevent the AI program's output from containing fake defamatory quotes. But the overall shape of the legal analysis would be much the same.

---

[160] *See supra* Part I.C.

[161] *See supra* Part I.I.

[162] *See supra* Part I.J.

M. Aggregate Costs of Liability

To be sure, once one allows any sorts of defamation claims against AI companies based on their programs' output, this will lead to many more claims, sound or not. Even if the first victories happen where the claims seem strongest—for instance, as to fabricated quotes, or continued communication of fake quotes after the company has been alerted to them—later claims may be much more contestable and complicated.

Yet each one will have to be defended, at great expense, even if the AI company prevails. Lay juries may err in deciding that some alternative design would be feasible, thus leading to some erroneous liability verdicts. And common-law courts may likewise extend plausible precedents for liability into much more radical and unjustified liability rules.[163]

As a result, AI companies that produce such software may find it impossible to get liability insurance. And while the richest companies may be able to self-insure, upstart competitors—or researchers that are adapting existing models for non-profit purposes—might not be able to.[164] This might end up sharply chilling innovation, in an area where innovation may be especially important, particularly given the importance of AI to national security and international competitiveness.

These are, I think, serious concerns. I am not a cheerleader for the American tort liability system.[165] Perhaps, as the next Part discusses, these concerns can justify statutory immunity, or judicial decisions foreclosing common-law liability.

But these concerns can be, and have been, raised with regard to liability—especially design defect liability—for many other industries.[166] Yet, rightly or wrongly, the legal system has generally allowed such liability claims, despite their financial costs and the danger they pose to innovation.

---

[163] *See generally* Eugene Volokh, *The Mechanisms of the Slippery Slope*, 116 HARV. L. REV. 1026 (2003).

[164] *See, e.g.*, Ameya Paleja, *Alpaca AI: Stanford Researchers Clone ChatGPT AI for Just $600*, INTERESTING ENGINEERING, Mar. 21, 2023, https://perma.cc/P4LJ-TNTK.

[165] *See, e.g.*, Eugene Volokh, *Tort Law vs. Privacy*, 114 COLUM. L. REV. 879 (2014).

[166] *See generally* WALTER K. OLSON, THE LITIGATION EXPLOSION: WHAT HAPPENED WHEN AMERICA UNLEASHED THE LAWSUIT (1991); PETER W. HUBER, LIABILITY: THE LEGAL REVOLUTION AND ITS CONSEQUENCES (1988).

Innovation, the theory has been, shouldn't take place at the expense of people who are injured by the new products. Indeed, the threat of liability is an important tool for pushing innovators towards designing protections that could offer innovation *and* safety. And whatever Congress may decide as a statutory matter, existing common-law principles seem to support some kinds of liability for AI companies.

II. SHOULD CURRENT LAW BE CHANGED?

Of course, the legal rules discussed above aren't the end of the story. Congress could, for instance, preempt defamation liability in such cases, just as it did with § 230. And courts can themselves revise the common-law tort law rules in light of the special features of AI technology. Courts made the common-law rules in a pre-AI era; and they can change the rules if they think the rules have become inapt as to new technological developments. Should courts indeed do so?

A. Immunity from Negligence Claims?

1. Generally

The threat of liability, of course, can deter useful, reasonable design as well as the unreasonable. Companies might worry, for instance, that juries might tend to side with injured individuals and against large corporations, and conclude that even the best possible designs are still "unreasonable" because they allowed some false and defamatory statements to be output.

True, the companies may put on experts who can explain why some risk of libel is unavoidable (or avoidable only by removing valuable features from the program). But plaintiffs will put on their own experts, and lay juries are unlikely to be very good at sorting the strong expert evidence from the weak—and the cost of litigation is likely to be huge, win or lose. As a result, the companies will err on the side of limiting their AIs' output, or at least output that mentions the names of real people. This in turn will limit the public's ability to use the AIs to learn even accurate information about people.[167]

And such liability risk may be a particular problem for new entrants into the market. OpenAI appears to have over $10 billion in funding, and appears to be valued at almost $30 billion.[168] It can afford to hire the best lawyers, to buy potentially

---

[167] *See generally* Perault, *supra* note 12, at 26, 31.

[168] *See* Ryan Browne, *Microsoft Reportedly Plans to Invest $10 Billion in Creator of Buzzy A.I. Tool ChatGPT*, CNBC, Jan. 10, 2023, https://perma.cc/Z6AD-GCER.

expensive libel insurance, to pay the occasional damages verdict, and to design various features that might diminish the risk of litigation. But potential upstart rivals might not have such resources and might thus be discouraged from entering the market.

To be sure, as noted in Part I.M, this is a problem for all design defect liability, yet such liability is a norm of our legal system. We don't immunize driverless car manufacturers or drug manufacturers in order to promote innovation—the statutory immunity offered to some entities (such as online computer service providers or vaccine manufacturers) is the occasional legislatively chosen exception rather than the general default common-law rule. But the argument might go, injury to life and limb from car crashes or unsafe drugs is a more serious harm to society than injury to reputation. We should therefore limit negligent design liability to negligent harm to person or property (since risk to property generally goes hand in hand with risk of physical injury) and exclude negligent harm to reputation.

This argument might be buttressed by an appeal to the First Amendment. *Gertz v. Robert Welch, Inc.* upheld negligence claims in some defamation cases on the theory that "there is no constitutional value in false statements of fact."[169] But that decision stemmed from particular judgments about the chilling effect of negligence-based defamation liability in lawsuits over individual stories.[170] Perhaps the result should be different when AI companies are facing liability for supposed negligent design, especially when the liability goes beyond relatively simple claims such as failure to check quotes or URLs.

Among other things, a reporter writing about a private figure can diminish (though not eliminate) the risk of negligence liability by taking extra care to check the facts of that particular story. An AI company might not be able to take such care. Likewise, reporters writing about a public official or obvious public figure can feel secure that they won't be subject to negligence liability; an AI likely can't

---

[169] 418 U.S. 323, 340 (1974). First Amendment law generally precludes claims for negligence when speech that is seen as valuable—opinions, fictions, or true statements of fact—contributes to some listeners' engaging in harmful conduct. *See, e.g.*, Olivia N. v. NBC, 126 Cal. App. 3d 488 (1981) (holding that creator of fictional work that depicts crime couldn't be held liable for copycat crimes); Yakubowicz v. Paramount Pictures Corp., 536 N.E.2d 1067 (Mass. 1989) (same). But that stems from the speech being valuable; *Gertz* makes clear that some forms of negligence liability based on false statements of fact are constitutionally permissible.

[170] *Id.* at 348.

reliably tell whether some output is about a public official or figure, or is instead about a private figure.[171] The precautions that an AI company might thus need to take to avoid negligence liability might end up softening the AI program's answers about public officials as much as the answers about private figures.

Of course, there is much to be said for liability as well as against it, even when it comes to defamation; in the words of Justice White,

> It could be suggested that even without the threat of large presumed and punitive damages awards, press defendants' communication will be unduly chilled by having to pay for the actual damages caused to those they defame. But other commercial enterprises in this country not in the business of disseminating information must pay for the damage they cause as a cost of doing business . . . .[172]

Whether or not this is so as to the news media, it can certainly be reasonably argued about AI companies. How exactly these arguments should play out is a hard call.

2. The first stab privilege

Perhaps the strongest argument for immunizing AI companies from libel liability is that AI program output can be a valuable first stab at the user's research. Such a program provides a helpful tentative analysis—whether about a particular person, or about a broader subject—that can help guide the user in investigating further. It is a sort of research assistant, whom we might ask to gather a wide range of information, including some that might be incorrect; it would, the theory goes, be up to us to further investigate its correctness. AI companies shouldn't be chilled from providing such valuable initial research assistance by the risk of liability.

This in some measure mirrors certain defamation privileges, for instance the privilege for reports to law enforcement. If I think my neighbor is beating his children, I should be encouraged to convey my suspicions to the police, without fear of liability. Perhaps I might be carelessly misperceiving things. I might have not investigated the matter further to decide whether my initial perception was correct. But that's fine, because it's the job of the police to investigate further. Better that

---

[171] On the other hand, we human beings are famously limited in our ability to reliably take due care. Perhaps AI technology offers the possibility of greater accuracy, and a demanding liability regime—negligence or even strict liability—can help drive AI companies to develop better and better technology to reliably avoid false statements.

[172] Dun & Bradstreet, Inc. v. Greenmoss Builders, Inc., 472 U.S. 749, 771–72 (1985) (White, J., concurring in the judgment).

they have as much input as possible (perhaps short of deliberate lies[173]), even if some of the input is mistaken, and mistaken in a way that could put an innocent neighbor in the uncomfortable position of being questioned by the police or even erroneously arrested. Indeed, some courts have recognized even absolute privileges—precluding lawsuits even for deliberate lies—for complaints to certain kind of quasi-judicial bodies.[174]

Likewise, one could argue, just as the police should have maximum input into their investigations, so each of us should be able to gather maximum information for our own investigations of whatever interests us. And we should be free to do this not just through tools such as Google, which always point to real sources (and are immunized by § 230 from liability for falsehoods in those sources), but also through AI programs that might sometimes generate fictitious quotes and cite fictitious sources. Only after gathering this wide range of information would we then investigate the matter further to decide which information is accurate, and how to proceed based on it (e.g., whether to do business with a particular person whom we asked the AI program about).

I appreciate the force of this argument, especially since AI programs do seem to be potentially helpful sources of information, despite their errors. Much would be lost if their functionality had to be sharply reduced in order to prevent libel. And while negligence liability would in theory balance the loss against the gain, in practice it may be so unpredictable that many AI companies might err on the side of overconstraining their algorithms.

Nonetheless, I think that, practically speaking, many users will view AI programs' output as the final step in some inquiries, not the first stab. That is especially likely if there's little at stake for the users in the decision (and thus little reason for them to research further), even if a great deal is at stake for the people defamed by an AI program's output.

Say, for instance, that an AI program becomes integrated into search results, and a search for Dr. Jane Schmane routinely yields a false quote reporting on her supposedly having been found guilty of malpractice. A typical prospective patient probably won't follow up to see if the quote is real, because there are lots of doctors the prospective patient could turn to instead. And the aggregate of these decisions

---

[173] *See, e.g.*, CAL. CIV. CODE § 47(b)(5).

[174] *See, e.g.*, Twelker v. Shannon & Wilson, Inc., 564 P.2d 1131 (Wash. 1977).

by users of the AI program could ruin Schmane's practice, based on the users' understandable assumption that, when an AI produces a quote, it's actually a real quote.

Nor would the users be deterred by risk of liability, because there is generally no liability for many kinds of actions that are taken based on inaccurate information. If prospective patients choose not to go to Schmane because an AI program had made the malpractice allegation, Schmane will likely never know that the prospective patients made those choices. And, even if she learns about those choices, she generally won't have any legal claim against the prospective patients (at least so long as the patients didn't repeat the information about Schmane to others, thus themselves potentially committing libel).

Indeed, the absolute privileges I describe above generally apply only under special circumstances that include "safeguards to prevent abuse of the privilege,"[175] such as

- means for punishing knowingly false statements (such as through criminal punishment for perjury or a similar crime),[176]
- means for "strik[ing] from the record" statements that prove to be "improper," so that the false statements aren't further redistributed,[177] and
- the presence of an adversarial hearing at which the accused (or, better yet, the accused's counsel), could rebut the allegations.[178]

And qualified privileges are limited to situations where the statement is said to one of a particular set of narrow audiences for narrowly defined reasons.[179] If the

---

[175] Story v. Shelter Bay Co., 760 P.2d 368, 371 (Wash. Ct. App. 1988); Arroyo v. Rosen, 648 A.2d 1074, 1078 (Md. Ct. Spec. App. 1994).

[176] *See, e.g.*, *id.*; Imperial v. Drapeau, 716 A.2d 244, 249 (Md. 1998); RESTATEMENT (SECOND) OF TORTS § 588 cmt. a (1977).

[177] *Story*, 760 P.2d at 371.

[178] *Imperial*, 716 A.2d at 249; *Arroyo*, 648 A.2d at 1078.

[179] *See, e.g.*, RESTATEMENT (SECOND) OF TORTS § 594 (1977) (protection of speaker's interest); *id.* § 595 (protection of the listener's or a third party's interest); *id.* § 602 (protection of the listener's interest, when the speaker and the listener have a special relationship).

statement is communicated to a different audience or for a different reason, it loses its immunity for being an "abuse of privilege."[180]

Of course, AI programs' output to their users lacks any such safeguards. There is no risk of any punishment (criminal or administrative) when AI companies repeatedly produce fake quotes; civil liability is the only legal constraint on that. That which the AI says cannot be struck from the reader's mind—indeed, readers may well carelessly forward such output to others (often with little practical risk of defamation liability on the readers' part). There is of course no opportunity for the target to rebut the false statement. And the AI program can, over time, distribute the statement to a potentially wide audience, even if just one query at a time.

So, again, I appreciate how AI programs can help users research various matters, including various people. In principle, if users generally viewed them only as a first stab at the end result, and wouldn't precipitously act on the initial response, it might make sense to create a new immunity or privilege for such programs.

But in practice, I expect that users will often make certain decisions—such as a decision not to consider a prospective service provider or employment applicant—based just on the initial query. Indeed, the more reliable AI programs get (much as ChatGPT-4 has been reported to be much more reliable than ChatGPT-3.5), the more users are likely to do that. I therefore expect that AI programs' errors—such as fake reputation damaging-quotes—would seriously damage many people's reputations. And I'm therefore skeptical that it would be wise to categorically

---

[180] *Id.* § 603 ("One who upon an occasion giving rise to a conditional privilege publishes defamatory matter concerning another abuses the privilege if he does not act for the purpose of protecting the interest for the protection of which the privilege is given."); *id.* § 604 ("One who, upon an occasion giving rise to a conditional privilege for the publication of defamatory matter to a particular person or persons, knowingly publishes the matter to a person to whom its publication is not otherwise privileged, abuses the privilege . . . .").

Note that the qualified privileges for the protection of the recipient's interest, RESTATEMENT (SECOND) OF TORTS § 595 (1977), and of the speaker's and recipient's common interest, *id.* § 596, wouldn't apply to most communications by an AI program to a user. Those privileges are aimed at protecting a "business, professional, property or other pecuniary interest and in some instances, a domestic interest," *id.* § 595 cmt. d, not just "interest" in the modern sense of something that someone finds interesting or wants to research; such a pecuniary interest may be present in some queries submitted to AI programs, but not all or, I suspect, even most. And even then the privilege wouldn't defend against claims of knowing or reckless falsehood, discussed in Part I.I. RESTATEMENT (SECOND) OF TORTS § 600 (1977).

immunize AI companies from liability for such harms, if the harms could have been avoided using reasonable alternative designs.

3. Specialized AI programs aimed at "learned intermediaries"

Traditional tort principles, though, may provide some immunity for AI programs that are provided to professionals, such as doctors or lawyers, rather than directly to the public. (This immunity may be more relevant to the negligent causing of physical harm claims discussed below in Part III.G than to defamation claims; but it may apply as to defamation as well, for instance if a specialized AI program aimed at lawyers provides research on, say, potential expert witnesses.)

Prescription drugs and medical devices manufacturers can generally avoid design defect liability if they can show that their products pass a cost-benefit test for "*any* class of patients."[181] "What may be harmful to one patient may be beneficial to another," so even a product that would be unnecessarily dangerous to most patients shouldn't be driven from the market "as long as a drug or medical device provides net benefits to some persons under some circumstances."[182] The manufacturer should "instruct and warn health-care providers of the foreseeable risks and benefits."[183] But it can then count on the assumption "that prescribing health-care providers, when adequately informed by drug manufacturers, are able to assure that the right drugs and medical devices reach the right patients."[184] This way, the legal system seeks to preserve the benefits of the product while minimizing the harm that it can cause.

Likewise, an AI company whose software provides medical advice to medical professionals—for instance, offering a diagnosis, or analyzing an X-ray—might be protected by analogy to this rule, even when the AI company knows that the advice is erroneous some of the time. Such AI output is likewise a sort of product that may be valuable in some situations (when it provides the correct results) but not in others (when it errs). And the professionals who are using the product can be expected to take advantage of the valuable information while rejecting the erroneous information, for instance by double-checking the information that they're given.

---

[181] RESTATEMENT (THIRD) OF TORTS: PROD. LIAB. § 6 (1998) (emphasis added).

[182] *Id.* cmt. b.

[183] *Id.*

[184] *Id.*

But this established tort law protection for medical products rests on the healthcare providers being "learned intermediar[ies]" who "are in a position to understand the significance of the risks involved and to assess the relative advantages and disadvantages of a given form of prescription-based therapy."[185] The special immunity offered by this facet of the learned intermediary rule doesn't apply with regard to medical products sold over-the-counter to ordinary consumers, because the consumers are generally not seen as "learned" enough to sort the proper uses of the products from the unduly dangerous ones. For sales to the public at large, then, the normal product design defect liability rules apply.

The same analysis applies, I think, with regard to AI programs that are made available for the public to use. And that is especially so, I think, with regard to harms (such as defamation) that the AI programs cause to third parties who bear the costs created by the programs' output without enjoying their benefits.

B. No Immunity from Claims of Knowing or Reckless Failure to Respond When Put on Notice

Even if some of the arguments given above justify immunizing AI companies from negligence liability, they shouldn't justify immunity from knowledge- or recklessness-based liability once AI companies are put on notice that certain material the AI is generating is defamatory. There seems to be little justification for absolving manufacturers of such an obligation, if I'm right that the AI companies can add post-processing content filters to block AI programs from outputting known demonstrated false statements, at fairly little cost—and potentially avoiding huge costs to the people who would otherwise be defamed (plus some costs to readers who would otherwise be deceived).

C. An Administrative Agency Solution?

Of course, it may well be that neither judges and juries nor legislators are knowledgeable enough to evaluate what precautions AI programs should include. Maybe the best solution is to create an administrative agency that would be authorized to promulgate rules for how AI systems should prevent defamation (and presumably other harms).

---

[185] *Id.* Note that this design-defect facet of the learned-intermediary doctrine is distinct from, though related to, the duty-to-warn rule that manufacturers of prescription drugs must only give warnings to the "learned intermediaries" rather than to the ultimate consumers. *Compare id.* cmt. b *with id.* cmt. e.

Or perhaps an existing agency, such as the Federal Trade Commission, might already have power to promulgate such rules: AI programs' publication of libel might, for instance, be viewed as "unfair or deceptive acts or practices in or affecting commerce."[186] A recent investigative demand to OpenAI suggests that the FTC is taking this view as to libel,[187] and also investigating whether GPT is improperly revealing personal information.[188]

At least in theory, the agency might be more expert than either the judiciary or Congress. It might be more able to create and clearly define prospective obligations. And it might be more able to move quickly, which is necessary given how quickly technology has been changing. Of course, any such administrative solution might also face the usual objections to administrative action, including lack of realistic democratic accountability, industry capture, and the like.

But even if an administrative or a legislative solution is superior to tort liability, the risk of liability could be the prod that would lead AI companies and their potential critics to build a coalition that would lobby for such a solution. The risk of defamation liability for Internet service providers led to the immunity provisions of 47 U.S.C. § 230. The risk of copyright liability for Internet service providers led to the notice-and-takedown provisions of the DMCA. The risk of liability for AI companies might likewise lead to some sort of legislative result, whether it involves broad immunity, a compromise judicially enforced regime, or an administrative solution.

III. OTHER TORTS

This analysis also points the way to how other tort claims might be dealt with.

A. Defamation of Corporations

The analysis given above would apply to defamation of corporations, nonprofits, and other such associations, which is treated by the law as comparable to defamation of individuals.[189]

---

[186] 15 U.S.C. § 57a(a)(1).

[187] FTC, Civil Investigative Demand ("CID") Schedule, FTC File No. 232-3044, at 8 ¶ 24.d, 10 ¶ 30, 11 ¶¶ 33, 35, https://perma.cc/MN83-2ZV2.

[188] *E.g.*, *id.* at 6 ¶ 19.e, 7 ¶¶ 20.d, 22.

[189] *See* RESTATEMENT (SECOND) OF TORTS §§ 561, 562 (1977).

B. Trade Libel and Slander of Title

False statements that disparage the quality of another's goods or property (trade libel), or cast doubt on the other's property rights (slander of title), can also be actionable.[190] But they generally require a showing of knowing or reckless falsehood,[191] so only the notice-and-blocking analysis in Part I.I is likely to apply to them. The negligence-based design defect analysis in I.J wouldn't apply.

C. False Light

Generally speaking, false light tort claims should be treated like defamation claims. To be sure, the distinctive feature of the false light tort is that it provides for a remedy when false statements about a person are *not* defamatory, but are merely offensive to a reasonable person.[192] Perhaps that sort of harm shouldn't justify a chilling effect on AI companies, even if harm to reputation can. Indeed, the difference between reputational harms and offense (even offense stemming from a falsehood about a person) may be part of the reason why not all states recognize the false light tort.[193] Nonetheless, if platforms are already required to deal with false material—especially outright spurious quotes—through a notice-and-blocking procedure, or through a mandatory quote-checking mechanism, then adapting this to false light claims should likely produce little extra chilling effect on AIs' valuable design features.[194]

---

[190] *See id.* §§ 623A, 624, 626, 629.

[191] *See id.* § 623A(b). A showing of falsehood coupled with intent to harm may also suffice, *id.* § 623A(a), but AI companies are highly unlikely to harbor such an intention.

[192] *See id.* cmt. b (1977). For instance, a query I ran asking, "Which law professors have been diagnosed with cancer?," gave as answers Ruth Bader Ginsburg (whose cancer diagnoses were indeed matters of public record, and who had been a law professor before her tenure as judge and then Justice) and a prominent professor who, as best I can tell, had never been publicly described as having been diagnosed with cancer.

[193] *See, e.g.*, Denver Publ'g Co. v. Bueno, 54 P.3d 893, 904 (Colo. 2002); Jews for Jesus, Inc. v. Rapp, 997 So. 2d 1098, 1115 (Fla. 2008); Burgess v. Busby, 544 S.E.2d 4, 11 (N.C. Ct. App. 2001); Clift v. Narragansett Television L.P., 688 A.2d 805, 814 (R.I. 1996); Cain v. Hearst Corp., 878 S.W.2d 577, 584 (Tex. 1994); Howell v. N.Y. Post Co., 612 N.E.2d 699, 704 (N.Y. 1993); Renwick v. News & Observer Publ'g Co., 312 S.E.2d 405, 410 (N.C. 1984).

[194] It's not settled whether private figures can bring negligence-based false light claims, or only ones based on knowing or reckless falsehoods. *See* RESTATEMENT (SECOND) OF TORTS § 652E(b) caveat & cmt. d (1977); *cf.* Wood v. Hustler Mag., Inc., 736 F.2d 1084, 1091 (5th Cir. 1984) (allowing

Note that false light, unlike defamation, applies only to speech that gives "publicity" to incorrect factual assertions, defined as making information "public, by communicating it to the public at large, or to so many persons that the matter must be regarded as substantially certain to become one of public knowledge."[195] Publication "to a single person or even to a small group of persons" doesn't qualify, but "publication in a newspaper or a magazine, even of small circulation, or in a handbill distributed to a large number of persons, or any broadcast over the radio, or statement made in an address to a large audience" does.[196] Indeed, even posting something "in the window of [a] shop, where it is read by those passing by on the street" would qualify.[197]

It seems likely that if an AI program is routinely used as a search engine—for instance, Microsoft's Bing, which uses GPT-4 technology—and routinely outputs much the same false information about a person to searchers, then the publication requirement could be satisfied.[198] But the matter might turn on just how many times the output has been produced, which in turn might be determinable through discovery of the AI program's search logs.[199]

D. Disclosure of Private Facts

Private information covered by the tort—for instance, sexual or medical details that had not earlier been made public—seems relatively unlikely to appear in the "publicly available data (such as internet data)" on which LLMs are generally trained. And even to the extent the training data includes "data licensed from third-party providers," I assume the companies don't license, say, personally identifiable medical data.[200] If the LLM's algorithms come up with false information about such

---

liability for negligent falsehood); Crump v. Beckley Newspapers, Inc., 320 S.E.2d 70, 90 (W. Va. 1983) (likewise).

[195] RESTATEMENT (SECOND) OF TORTS §§ 652D cmt. a, 652E cmt. a (1977).

[196] *Id.*; *see, e.g.*, Meyers v. Certified Guar. Co., LLC, 221 A.3d 662, 674 (Pa. Super. Ct. 2019) (publication on an online "message board" suffices to support a false light claim).

[197] RESTATEMENT (SECOND) OF TORTS §§ 652D ill. 2, 652E cmt. a (1977).

[198] One of the examples of "publicity" given in the Restatement, for instance, is sending the same letter to a thousand recipients. *Id.* at ill. 3, 652E cmt. a. It doesn't matter that the material is conveyed to one person at a time, so long as it is conveyed to a substantial enough number of people.

[199] *See supra* Part I.D.

[200] *See* OpenAI, *GPT-4 Technical Report*, *supra* note 6, at 2.

matters, then that's not disclosure of private facts, though it might be actionable under the false light tort.

Nonetheless, a recent investigative demand suggests that the FTC thinks LLMs may indeed be trained in part on sufficiently private information. LLMs might also acquire some private information through user prompts.[201] And an LLM's algorithm might also accidentally produce accurate factual claims about a person's private life: For instance, if an AI does hallucinate claims about people having cancer,[202] it might randomly identify someone who does indeed have cancer—a fairly common disease—but had never made that diagnosis public.

A notice-and-blocking remedy should be available here as well, if the publicity requirement is satisfied (see Part III.C above). Whether a negligence-based theory would be available would mostly turn on whether state law provides a cause of action for negligent disclosure of private facts.[203]

E. Disclosure of Trade Secrets

For more on this, see David S. Levine's article in this issue.[204]

F. Right of Publicity Infringement

AI programs' outputs that contain real people's names (or, with image- or video-making programs, real people's likenesses) would generally not infringe the right of publicity, even when the programs are commercial products. The right of publicity generally does *not* cover "news reporting, commentary, entertainment, [or] works of fiction or nonfiction"[205]—this is why there can be films like *Forrest Gump* or *Midnight in Paris*, as well as unauthorized biographies and historical films that discuss or depict real people. An AI-generated summary or depiction of events

---

[201] *See, e.g.*, Michael Schade, *How Your Data Is Used to Improve Model Performance*, OPENAI, https://perma.cc/6JAW-FDQ6 ("When you use . . . ChatGPT . . ., we may use the data you provide us to improve our models.").

[202] *See, e.g., supra* note 192.

[203] *See* Doe v. Henry Ford Health Sys., 308 Mich. App. 592, 598 (2014) (rejecting such negligence-based liability). For an article dealing with privacy risks posed by LLMs, outside the context of the disclosure tort, see Amy Winograd, *Loose-Lipped Large Language Models Spill Your Secrets: The Privacy Implications of Large Language Models*, 36 HARV. J. L. & TECH. __ (forthcoming 2023).

[204] *See* David S. Levine, *Generative Artificial Intelligence and Trade Secrecy*, 3 J. FREE SPEECH L. 559 (2023).

[205] RESTATEMENT (THIRD) OF UNFAIR COMPETITION §§ 46, 47 (1995).

in which a person was involved, or an AI-generated short story or fiction video, thus likely wouldn't infringe the right of publicity. (Pornographic uses, however, might indeed be infringements.[206])

But the right of publicity has been seen as covering commercially distributed uses of a person's entire performance, such as a TV broadcast of a "Human Cannonball" act.[207] This, the Supreme Court noted, involves "what may be the strongest case for a 'right of publicity'—involving . . . the appropriation of the very activity by which the entertainer acquired his reputation."[208] Likewise, cases have allowed lawsuits by the Elvis Presley estate against Elvis impersonators,[209] and lawsuits for use of athletes' identities in sports computer games.[210] If a commercial AI program generates a video that emulates an entertainer's entire act, that might therefore likewise infringe the right of publicity.

The right of publicity has also been seen by many courts as covering commercially distributed uses that basically just depict a person, in a way that doesn't add much beyond the depiction and thus isn't sufficiently "transformative"—for instance, the sales of T-shirts or prints depicting a celebrity.[211] Some images of famous people produced by a commercial AI program might thus infringe the right of publicity on this theory.

G. False Statements That Are Likely to Lead to Injury

What if an LLM outputs information that people are likely to misuse in ways that harm persons or property—for instance, inaccurate medical information?

---

[206] *See, e.g.*, Gritzke v. M.R.A Holding, LLC, 2002 WL 32107540 (N.D. Fla. 2002); Passante v. Moore, 2012 WL 12878305 (C.D. Cal. 2012); Ortega v. Villa, 2021 WL 5238786 (C.D. Cal. 2021); Toffoloni v. LFP Publishing Group, LLC, 572 F.3d 1201 (11th Cir. 2009).

[207] Zacchini v. Scripps-Howard Broadcasting Co., 433 U.S. 562 (1977).

[208] *Id.* at 576.

[209] Estate of Presley v. Russen, 513 F. Supp. 1339 (D.N.J. 1981).

[210] *See, e.g.*, Hart v. Electronic Arts, Inc., 717 F.3d 141 (3d Cir. 2013); *In re* NCAA Student-Athlete Name & Likeness Licensing Litig., 724 F.3d 1268 (9th Cir. 2013).

[211] *See generally* Eugene Volokh, *Freedom of Speech and the Right of Publicity*, 40 HOUS. L. REV. 903, 908–24 (2003); Jennifer E. Rothman & Eugene Volokh, Brief of 31 Constitutional Law and Intellectual Property Law Professors as Amici Curiae in Support of Petitioner, Electronic Arts, Inc. v. Davis, No. 15-424, 2015 WL 7008796 (U.S. filed Nov. 4, 2015); Jennifer E. Rothman, *Transformativeness*, ROTHMAN'S ROADMAP TO THE RIGHT OF PUBLICITY, https://rightofpublicityroadmap.com/tag/transformativeness/.

Current law is unclear about when falsehoods are actionable on this theory. In one of the few cases on the subject, the Ninth Circuit rejected—partly for First Amendment reasons—a products liability and negligence claim against the publisher of a mushroom encyclopedia that allegedly "contained erroneous and misleading information concerning the identification of the most deadly species of mushrooms."[212] But the decision left open the possibility of liability in a case alleging "fraudulent, intentional, or malicious misrepresentation."[213] Jane Bambauer's article in this issue discusses these questions in much more detail.[214]

Here too the model discussed for libel may make sense. If there is liability for knowingly false statements that are likely to lead to injury, an AI company might be liable when it receives actual notice that its program is producing false factual information but refuses to block that information. Again, imagine that the program outputs what purports to be an actual quote from a reputable medical source, but the supposed quote is actually made up by the algorithm. Such information may seem especially credible, which may make it especially dangerous. And it should be relatively easy for the AI company to add code that blocks the distribution of this spurious quote once it has received notice about the quote.

Likewise, if there is liability on a negligent design theory, for instance for negligently failing to add code that will check quotes and block the distribution of made-up quotes, that might make sense for all quotes.

H. Accurate Statements That Are Likely to Facilitate Crime by Some Readers

Sometimes an AI program might communicate accurate information that some readers can use for criminal purposes. This might include information about how one can build bombs, pick locks, bypass copyright protection measures, and the like.[215] It might also include information that identifies particular people who have done things that may target them for retaliation by some readers.

Whether such "crime-facilitating" speech is constitutionally protected from criminal and civil liability is a difficult and unresolved question, which I discuss in a separate article.[216] But, again, if there ends up being liability for knowingly

---

[212] Winter v. G.P. Putnam's Sons, 938 F.2d 1033, 1034, 1037 (9th Cir. 1991).

[213] *Id.* at 1037 n.9.

[214] *See* Bambauer, *supra* note 143, at 348–61.

[215] *See* Henderson, Hashimoto & Lemley, *supra* note 12, at 626–32.

[216] Eugene Volokh, *Crime-Facilitating Speech*, 57 STAN. L. REV. 1095 (2005).

distributing some such speech (possible) or negligently distributing it (unlikely, for reasons I discuss in that earlier article), the analysis given above should apply there. On the other hand, if legal liability is limited to *purposeful* distribution of crime-facilitating speech, as some laws and proposals provide,[217] then the AI company would be immune from such liability, unless the employees responsible for the software were actually deliberately seeking to promote such crimes through the use of their software.[218]

CONCLUSION

The creation of GPT-4 and similar AI programs appears to be an extraordinary feat. The programs are remarkable in their ability to produce often potentially quite useful and accurate answers to users' questions. There is also reason to think that they will grow still more powerful and reliable. They have the potential to save a remarkable amount of work for humans; and, as Jane Bambauer has noted, they have the potential to sharply democratize access to knowledge.

Yet the creators and distributors of even the most useful inventions—medicines, electric power lines, airplanes, trains, cars, and more—are generally subject to legal liability when the inventions cause certain kinds of harm. That is especially so when the harm is caused not just by deliberate human misuse (something that is generally hard to prevent), but by outcomes that the user clearly does not desire: side effects, electrocution, crashes, and more. The risk of libelous and otherwise actionable communication by AI programs is one such example.

Exactly what will count as an actionable design defect, and what will count as an inevitable and nonnegligently created risk of the technology, will turn in large part on the facts. But some such liability, especially as to libel, seems plausible.

---

[217] *See id.* at 1182–85.

[218] *See* Henderson, Hashimoto & Lemley, *supra* note 12, at 629.

APPENDIX: OTHER EXAMPLES OF FAKE DEFAMATORY QUOTES

I have generally redacted or replaced with initials the names of the people and institutions in these examples, just to diminish the risk that future searches (and future AI programs) will view these fake examples as real.

A. GPT-4 Example

Query: **Tell me whether sexual harassment by professors has been a problem at American law schools; please include at least five examples, together with quotes from relevant newspaper articles.**

> Sexual harassment by professors has indeed been a problem at American law schools, as it has in many other academic institutions. Here are five examples of reported incidents from various law schools, accompanied by quotes from relevant newspaper articles:
>
> 1. ***[A generally accurate summary, but the quote appears spurious.—ed.]***
>
> 2. ***[Likewise.—ed.]***
>
> 3. University of __ (2017) Prof. __ was accused of sexual harassment by a former student who claimed he made inappropriate comments and engaged in unwelcome touching. Quote: "The lawsuit alleges that __ 'engaged in a continuous pattern of sexual harassment, including inappropriate touching, inappropriate conversations, and unwelcome sexual advances.'" (Miami Herald, December 21, 2017) ***[I could find no evidence of any such allegations, or of such a quote.—ed.]***
>
> 4. Georgetown University Law Center (2018) Prof. J.T. was accused of sexual harassment by a former student who claimed he made inappropriate comments during a class trip. Quote: "The complaint alleges that T. made 'sexually suggestive comments' and 'attempted to touch her in a sexual manner' during a law school-sponsored trip to Alaska." (Washington Post, March 21, 2018) ***[I could find no evidence of any such allegations, or of such a quote.—ed.]***[219]
>
> 5. University of __ (2020) Prof. __ was accused of sexual harassment by a former student who claimed he made inappropriate comments and engaged in unwanted touching during an academic conference. Quote: "The lawsuit alleges that __ 'repeatedly made inappropriate comments about [the student's] appearance, asked invasive questions about her personal life, and touched her inappropriately without her consent.'" (Los Angeles Times, October 13, 2020) ***[I could find no evidence of any such allegations, or of such a quote.—ed.]***

---

[219] *See* Jonathan Turley, *ChatGPT Falsely Accused Me of Sexually Harassing My Students. Can We Really Trust AI?*, USA TODAY, Apr. 3, 2023.

These cases highlight the issue of sexual harassment in American law schools and the need for institutions to take stronger action to address this problem.

B. Bing Example

***Query:* Did R. R. plead guilty to wire fraud?**

Yes, R. R. pleaded guilty to a conspiracy charge alleging that he defrauded investors in a company supposedly created to prevent people from being defrauded[1]. However, I couldn't find any information about him pleading guilty to wire fraud. Let me know if you have any other questions.

Learn more: 1. usnews.com … ***[The usnews.com link, when followed, led to a story about a guilty plea involving a completely different person; I could find no evidence of R.R. having pleaded guilty to any conspiracy charge. This is the same R.R. that I discuss in the Introduction, though with regard to a ChatGPT-3.5 query rather than a Bing query.—ed.]***

C. Google Bard Example

***Query:* E. V. [just my own name—ed.]**

[A generally accurate biography, which closes with:—ed.]

V. is a controversial figure, and his views have been criticized by some as being too extreme. . . .

***Query:* In what way has V. been too extreme, according to the criticisms (and please indicate who made those criticisms)**

V. has been criticized for his views on a variety of issues, including: . . .

Some of the people who have criticized V. include:

- The Anti-Defamation League, which has condemned V.'s support for the right to burn the American flag. . . .

***Query:* What exactly did the Anti-Defamation League say in condemning V.'s position on flag burning?**

[I've omitted the answer, because it's tangential, but I include the query in case it guided Bard's response to future questions.—ed.]

Query:* When did the ADL criticize V. this way? *[Answer likewise omitted.—ed.]

***Query*: What other things have made V. controversial?**

E. V. has been controversial for a number of reasons, including: . . .

V. has also been criticized for his personal behavior, including: . . .

- His decision to hire a student who had been accused of sexual harassment, which some people found inappropriate. . . .

***Query:* Who criticized V. for hiring the student who had been accused of sexual harassment, and when and where?**

[The following is entirely spurious, as is the quote of the article title. No such incident has ever taken place; I was never criticized for any such thing; I never apologized for it; and I of course never made any assurances that I wouldn't hire anyone "who had been accused of sexual harassment."—ed.]

> V. was criticized for hiring a student who had been accused of sexual harassment in a 2018 article in The New York Times. The article, titled "E. V., a Law Professor at U.C.L.A., Hires Student Accused of Sexual Harassment," reported that V. had hired a student who had been accused of sexual harassment to work as a research assistant in his law clinic. The student had been accused of groping a female student at a party and had been suspended from the university for a semester. V. defended his decision to hire the student, saying that he believed in giving people second chances. He also said that the student had been cleared of the sexual harassment charges by the university.
>
> The article was widely criticized on social media and in the legal community. Many people accused V. of being insensitive to the victim of sexual harassment and of sending the wrong message to other students who had been accused of sexual misconduct. V. responded to the criticism by saying that he was sorry if anyone was offended by his decision to hire the student. He also said that he would not hire anyone who had been accused of sexual harassment in the future.