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]

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.

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*rey* 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.

---

[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.

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]

---

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.