AUTHORBOTS

Derek E. Bambauer & Mihai Surdeanu[*]

INTRODUCTION

ChatGPT has exploded into the popular consciousness in recent months, and the hype and concerns about the program have only grown louder with the release of GPT-4, a more powerful version of the software.[1] Its deployment, including with applications such as Microsoft Office,[2] has raised questions about whether the developers or distributors of code that includes ChatGPT, or similar generative pre-trained transformers, could face liability for tort claims such as defamation or false

[*] Professor of Law, University of Arizona James E. Rogers College of Law; Associate Professor of Computer Science, College of Science, University of Arizona. Authors are listed alphabetically. We owe thanks for helpful suggestions and discussion to Jane Bambauer, Saumya Debray, Dan Hunter, Gus Hurwitz, Alice Kwak, Irina Manta, Clay Morrison, Thinh Nguyen, Sergio Puig, Amy Stein, and Eugene Volokh. We thank Guy Forte for expert research assistance. We welcome comments at <derekbambauer@arizona.edu> and <msurdeanu@arizona.edu>.

[1] *See* Amelia Thomson-Deveaux & Curtis Yee, *ChatGPT Thinks Americans Are Excited About AI. Most Are Not.*, FIVETHIRTYEIGHT (Feb. 24, 2023), https://perma.cc/2PLE-R3MD; Drew Harwell & Nitasha Tiku, *GPT-4 Has Arrived. It Will Blow ChatGPT out of the Water.*, WASH. POST (Mar. 14, 2023).

[2] *See* Samantha Murphy Kelly, *Microsoft Is Bringing ChatGPT Technology to Word, Excel and Outlook*, CNN (Mar. 16, 2023).

light.[3] One important potential barrier to these claims is the immunity conferred by 47 U.S.C. § 230, popularly known as "Section 230."[4] In this Essay, we make two claims. First, Section 230 is likely to protect the creators, distributors, and hosts of online services that include ChatGPT in many cases. Users of those services, though, may be at greater legal risk than is commonly believed. Second, ChatGPT and its ilk make the analysis of the Section 230 safe harbor more complex, both substantively and procedurally. This is likely a negative consequence for the software's developers and hosts, since complexity in law tends to generate uncertainty, which in turn creates cost. Nonetheless, we contend that Section 230 has more of a role to play in legal questions about ChatGPT than most commentators do—including the principal legislative drafters of Section 230—and that this result is generally a desirable one.[5]

I. A FEW WORDS ABOUT TECHNOLOGY

A significant theme in popular media and even some scholarly commentary on ChatGPT is technopanic:[6] The software has achieved sentience,[7] or can craft code to "escape" from where it is hosted to other systems,[8] or will upend the e-commerce

---

[3] The Essay refers to ChatGPT throughout, since that is the focus of this set of papers, but its discussion applies to GPT code more broadly and, in most aspects, to the larger set of generative algorithms.

[4] *See generally* JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET (2019).

[5] *See* Cristiano Lima, *AI Chatbots Won't Enjoy Tech's Legal Shield, Section 230 Authors Say*, WASH. POST (Mar. 17, 2023).

[6] *See* Jaron Lanier, *There Is No A.I.*, NEW YORKER (Apr. 20, 2023); Eliezer Yudkowsky, *Pausing AI Developments Isn't Enough. We Need to Shut It Down*, TIME (Mar. 29, 2023) ("If someone builds a too-powerful AI, under present conditions, I expect that every single member of the human species and all biological life on Earth dies shortly thereafter.").

[7] *See Pause Giant AI Experiments: An Open Letter*, FUTURE OF LIFE INST. (Mar. 22, 2023), https://perma.cc/SF9F-G525.

[8] *See* Kristine Parks, *AI Expert Alarmed After ChatGPT Devises Plan to 'Escape': 'How Do We Contain It?'*, FOX NEWS (Mar. 20, 2023).

and online advertising industries,[9] or will corrupt children.[10] These fears are premature at best and counterproductive hysteria at worst. A few words about the ChatGPT technology can usefully clarify matters.

First, ChatGPT is not intelligent and is not approaching human reasoning capabilities. Anthropomorphizing software is both pervasive and harmful. ChatGPT's methods for "learning" information and producing responses to queries differ starkly from how people perform those tasks. People reason; GPTs regurgitate.[11] Thinking of ChatGPT as intelligent and self-aware is unhelpful for a variety of reasons; this Essay shows that one of those reasons is confusion in assessing how the software interacts with Section 230. This point about erroneously seeing sentience in software cannot be overemphasized. These tools are no more conscious or intelligent than Microsoft Clippy (or Bob) was, no matter how many articles outlets such as *WIRED* run incorrectly stating otherwise.[12] As this Essay discusses further in this section, this is a characteristically human tendency—to see humanity in nature and even inanimate objects. But cognitive errors, including ones driven by emotion, are poor grounds for policymaking or legal decisions. Most observers who subscribe to the more stark fears of artificial intelligence cannot formulate an empirical or positivistic reason for why they harbor such worries.[13] It's simply feelings. However, neither a stuffed animal nor an autocomplete tool is alive or sentient, even if we wish or fear that it were.[14] ChatGPT's style of crafting responses to our queries looks familiar, even human, but it manifestly is not. The discourse over the software would be immeasurably improved if it permanently placed this truth front and center.

Second, ChatGPT is, roughly, an auto-completion tool on steroids. It uses the information in its training data to build what it deems the most responsive answer to a user's query. The code is purely probabilistic: It builds a response that has the

---

[9] *See* Tripp Mickle, Cade Metz & Nico Grant, *The Chatbots Are Here, and the Internet Industry Is in a Tizzy*, N.Y. TIMES (Mar. 8, 2023).

[10] *See* Geoffrey A. Fowler, *Snapchat Tried to Make a Safe AI. It Chats with Me About Booze and Sex.*, WASH. POST (Mar. 14, 2023).

[11] *See* Noam Chomsky, *The False Promise of ChatGPT*, N.Y. TIMES (Mar. 8, 2023).

[12] *See* Will Knight, *Some Glimpse AGI in ChatGPT. Others Call It a Mirage*, WIRED (Apr. 10, 2023), https://perma.cc/LV4D-LEQZ.

[13] *See* Lanier, *supra* note 6.

[14] *Id.*

highest statistical likelihood of comprising what the user seeks, based on its training data and the associations the software computes between the words contained within it.[15] ChatGPT is, in other words, a pleaser. The software just plays a numbers game—it has no logical model of why certain words fit together, and no theoretical basis to evaluate the quality of its responses. This explains why ChatGPT provides responses that seem nonsensical or flatly wrong. Unlike Google or other search engines, ChatGPT is not an information retrieval tool—it is effectively a hallucination generation engine.[16] Computer scientists refer to ChatGPT's output as "beige bullshit": Beige because it is simply the response with the safest probability of being correct based upon the underlying training data, and bullshit because of both the hallucination problem and because ChatGPT is indifferent to truth.[17] The software's goal is to predict the next most likely word in a sequence. Whether that response is truthful is completely orthogonal to ChatGPT's design. By analogy: ChatGPT isn't playing chess; it is playing *Mad Libs*.[18] The software excels at guesswork based upon memorization (or, more accurately, assimilation of training data), but it has no capacity to reason or conceptualize.[19]

---

[15] OpenAI added a human feedback layer, where humans train GPT models to rank higher *complete* statements. *See* Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, NeurIPS 2022, at 3–4, 4 fig.2, https://perma.cc/Y4WX-UJD6 (describing this in the last two blocks of the figure). The goal of this component was to reduce the models' weighting of offensive or incorrect statements. However, since this component is proprietary and not open-source software, it is not clear what its contribution to the models' overall performance is.

[16] GPT-4 improves upon earlier versions regarding hallucinations, but OpenAI still cautions that the software "still is not fully reliable (it 'hallucinates' facts and makes reasoning errors)." *GPT-4*, OpenAI (Mar. 14, 2023), https://perma.cc/CSH5-GW95.

[17] *See* Harry G. Frankfurt, On Bullshit (2005) (defining "bullshit" as communication that is purely instrumental, intended to reach a particular goal without regard to the truth of the information conveyed).

[18] For those unfamiliar with the Mad Libs word puzzles, see Mad Libs, https://perma.cc/9EBK-RYL3; *Mad Libs*, Wikipedia, https://perma.cc/KW77-QNUT (last updated Nov. 29, 2022).

[19] *See* Lanier, *supra* note 6 ("[L]arge language model[s] like GPT-4 contain[] a cumulative record of how particular words coincide in the vast amounts of text that the program has processed. . . . When you enter a query consisting of certain words in a certain order, your entry is correlated with what's in the model; the results can come out a little differently each time, because of the complexity of correlating billions of entries."). Even Google's chief executive officer has made this mistake, and on 60 Minutes no less. *See* Pranav Dixit, *Researchers Accused Google and "60 Minutes" of Spreading AI "Disinformation,"* BuzzFeed News (Apr. 19, 2023), https://perma.cc/F67M-DEMP (discussing

Third, human cognitive biases explain much of the fear surrounding ChatGPT. People focus on the times that ChatGPT seems self-aware but downplay or forget the times that it is plainly spewing nonsense. Humans anthropomorphize things—we see faces in the clouds, and we see sentience in authorbots like ChatGPT. Relatedly, risk and loss aversion push us towards more frightening interpretations of the software's capabilities. It is much more interesting to read an article about the possibility that ChatGPT is on the verge of becoming HAL 9000 or Skynet than to read one pointing out that the software is basically a tool for creating crude rough drafts. The human tendency to view a computer-based interlocutor as sentient has a long history: It was first documented in the 1960s.[20] The hype surrounding ChatGPT has a similarly long provenance—Marvin Minsky, a pioneer in the field of artificial intelligence, claimed in 1970 that "In from three to eight years we will have a machine with the general intelligence of an average human being . . . . In a few months it will be at genius level, and a few months after that its powers will be incalculable."[21]

Fourth, ChatGPT's responses are highly dependent upon the queries it receives—again, the program is a pleaser. This can lead observers to mistake cause for effect. For example, the *New York Times* reporter who had a purportedly frightening experience with the bot failed to notice that ChatGPT began returning answers that seemed upset or emotional only once the reporter became upset, which was reflected in his queries.[22] ChatGPT is thus something of a mimic. Normatively undesirable outputs, such as false or defamatory ones, are more readily generated by inputs that push ChatGPT in that direction—in other words, by leading questions. Researchers have exploited this tendency to circumvent content restrictions and other security measures in generative algorithmic systems: They "us[e] carefully crafted and refined sentences . . . to exploit system weaknesses."[23] Defamatory

---

claim that Google's PaLM large language model had learned a new language without being trained on it; the language was, in fact, in Bard's training database).

[20] *See ELIZA*, WIKIPEDIA, https://perma.cc/DQ4W-B5E9 (last updated Apr. 17, 2023).

[21] Brad Darrach, *Meet Shaky, the First Electronic Person*, LIFE, Nov. 20, 1970, at 58D.

[22] *See* Kevin Roose, *A Conversation with Bing's Chatbot Left Me Deeply Unsettled*, N.Y. TIMES (Feb. 16, 2023); Tiernan Ray, *ChatGPT: What* The New York Times *and Others Are Getting Terribly Wrong About It*, ZDNET (Mar. 1, 2023), https://perma.cc/DTH8-5AHF.

[23] Matt Burgess, *The Hacking of ChatGPT Is Just Getting Started*, WIRED (Apr. 13, 2023), https://perma.cc/7ZB3-ZJDN.

blossoms often result from defamatory seeds. ChatGPT outputs errors—hallucinations—on its own, but it is also easily led astray based upon the queries with which it is prompted.

Fifth, ChatGPT creates responses that are a pastiche of the information in its training corpus. The software incrementally generates small text fragments that correspond to subparts of words called "tokens," which correspond roughly to syllables and other common sequences of characters. When ChatGPT receives a query, it remixes the data based upon these tokens, drawing on the underlying data as a whole. A sentence may have different tokens, from different sources, for each individual word. The mash-up character of a ChatGPT response makes it difficult to ascertain the provenance of the answer, which has important ramifications for Section 230. This also means that the underlying training data is critically important. Part of the reason that ChatGPT has been so successful with answering bar exam questions is likely that its training data includes published bar exams and bar exam preparation material—a phenomenon that machine learning researchers call "contamination."[24] There are relatively limited ways to frame test questions, at least on certain subjects, and the combination of a limited solution set and highly relevant training data makes the bar exam an easy test for ChatGPT to pass.[25]

Lastly, ChatGPT is also improving because OpenAI, its developer, has humans in the loop during the training process. Humans perform at least two functions in training ChatGPT.[26] First, people decide what new content to add to the training corpus, and when. Second, people curate responses—they rank or grade ChatGPT's output, teaching the software which outputs are more or less appropriate. Like the software's developers, its trainers might plausibly be seen as creating or developing, at least in part, ChatGPT's responses. Those trainers are either

---

[24] Unsurprisingly, ChatGPT is often successful at predicting a likely pattern of words in response to a query if the software has been trained on a close variant of the query and responses to it. *See supra* note 19. This mirrors how prospective lawyers study for the bar exam: training materials emphasize memorization over reasoning, particularly with multiple choice questions. Humans, however, lack ChatGPT's level of recall.

[25] OpenAI stated that they removed the evaluation sections of bar exams from GPT-4's training corpus. However, given the popularity and importance of bar exams, it is likely that such data has been widely replicated in multiple locations, beyond the official bar exam Web sites, that are included in the training data.

[26] *See* Ouyang et al., *supra* note 15.

OpenAI employees or contractors working for the firm; regardless, their interaction with the corpus might well place OpenAI outside Section 230's safe harbors.

Having humans in the loop for determining the content of ChatGPT training data and grading the software's responses raises at least three concerns. First, determinations of which output is better or worse, or more or less appropriate, depends significantly on the normative views of the humans making that judgment. We have no reason presently to doubt the good faith of ChatGPT's trainers, but an unscrupulous trainer could potentially insert false information into the training corpus to skew results. Second, human control over the training corpus creates the risk of contamination, either deliberate or inadvertent. For example, ChatGPT might perform well on state bar exams (required in most states to practice law) if trainers inserted previous bar exams, and potentially answers to them, into the training data. Lastly, OpenAI's conduct reveals well-known problems experienced by people who curate data and who are thereby exposed to problematic content that can lead to psychological harms.[27] OpenAI compounded the problem by outsourcing this emotionally difficult task to workers in Kenya whom it paid less than two dollars per hour.[28] Each of these concerns flows from the decision to have humans play a key role in training the ChatGPT software.

II. IS CHATGPT A SPEAKER FOR SECTION 230 PURPOSES?

The key question for the protection that Section 230 might afford ChatGPT is whether the bot is a speaker in that statute's framework. Section 230 effectively eliminates publisher liability online (and also distributor liability after *Zeran v. AOL*[29] and its progeny), leaving only speakers as potential "information content providers" under the statute. Thus, liability for claims such as defamation turns on whether ChatGPT is an information content provider responsible for the libelous content. Section 230 defines an information content provider (ICP) as "any person or entity that is responsible, in whole or in part, for the creation or development of

---

[27] *See* Miriah Steiger et al., *The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support*, 341 CHI '21, at 1 (May 7, 2021), https://perma.cc/7BVF-NY3D; Andrew Arsht & Daniel Etcovitch, *The Human Cost of Online Content Moderation*, JOLT DIGEST (Mar. 2, 2018), https://perma.cc/5Q5N-QPEC.

[28] *See* Billy Perrigo, *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic*, TIME (Jan. 18, 2023).

[29] Zeran v. America Online, Inc., 129 F.3d 327 (4th Cir. 1997).

information" made available via an interactive computer service. There seem to be four possibilities here.

First, and most simply, ChatGPT's mash-up of data in response to a query might have no basis in its training corpus. The bot may have combined unrelated information, made a mistake, or just produced a hallucination. In this case, ChatGPT is likely the ICP responsible, at least in part, for the defamatory content, and Section 230 does not apply. This possibility does have implications for defamation law, because the production of the inaccurate content does not result from human volition.[30] Moreover, even if one were to engage in the fiction that either OpenAI or the entity hosting the ChatGPT software were the speaker of the false information, it seems unlikely that the publication of that data could meet defamation doctrine's requirement of negligence or worse—measuring mental state is impossible when there is no mind.[31] Here, ChatGPT embodies the infinite monkey theorem: Given enough queries and time, the code will generate false content simply due to stochastic chance.[32] Imposing tort liability in this scenario is virtually useless from a deterrence perspective: A chatbot that makes any errors will, over time, eventually produce a false statement about a person in response to a query.

---

[30] Volitional conduct plays a complicated role in tort law, which typically focuses on the presence or absence of sufficient precautions (in negligence-based liability). Nonetheless, tort law generally requires "a causal connection between the (tortious) volitional conduct of the defendant and the harm the plaintiff has suffered." Richard M. Wright, *Substantive Corrective Justice*, 77 IOWA L. REV. 625, 639 (1992); *see also* Sheldon Nahmod, *Constitutional Damages and Corrective Justice: A Different View*, 76 VA. L. REV. 997, 1009 (1990) (noting that even in cases "where wrongdoing is grounded on negligence, it frequently turns out that no particular state of mind is required beyond volitional conduct"). We thank Jane Bambauer, Gus Hurwitz, and Eugene Volokh for helping elucidate this point.

[31] Plaintiffs have occasionally surmounted Section 230 immunity in software cases based on negligence theories. *See, e.g.*, Lemmon v. Snap, Inc., 995 F.3d 1085 (9th Cir. 2021). Negligent product design claims, if generally allowed to bypass Section 230, would eviscerate the statute's immunities. *Lemmon*, though, seems to result from the old wisdom that hard cases make bad law. Snapchat's Speed Filter allowed users to record their speed and superimpose it on photos or videos; users appeared to believe that Snapchat's rewards system would compensate them with virtual honors for capturing high-speed travel while running Speed Filter; and three people allegedly died under just those circumstances. *Id.* at 1088–89. And while the Ninth Circuit stated its analysis did "not [enable] a creative attempt to plead around the CDA," *id.* at 1094, that is exactly what it did.

[32] *See* Adam Frank, *The Infinite Monkey Theorem Comes To Life*, NPR (Dec. 10, 2013), https://perma.cc/J7NV-B2TW.

All software code of any appreciable length has bugs. Using defamation law to try to change that inevitability is to seek to command the tides.[33]

Second, ChatGPT's response might be based on information in its training data, albeit not word-for-word. Thus, the underlying claims can be found in the corpus, and ChatGPT reformulates them into its answer. The majority of Section 230 precedent treats this type of restating or summarizing as the traditional function of a publisher, which is distinct from an ICP and which is protected by Section 230.[34] The theory is that the defendant (here, ChatGPT) has not produced any new semantic content. Instead, it has merely repackaged existing content. So long as ChatGPT does not alter the underlying semiotics of the material in its training data, it is likely to enjoy immunity.

Third, ChatGPT's answer might reproduce, either precisely or nearly so, claims in its training data. It is possible that the bot might arrive at this response in roundabout fashion—perhaps it recreates the claim using its standard collage approach to data. In that case, a formalist court might consider ChatGPT as the ICP responsible for the answer. As a practical matter, though, independent creation will be nearly impossible to prove: ChatGPT does not track the provenance of its answers, and so there is no way to distinguish between the system regurgitating a pre-existing statement and creating a new one (in theory, ChatGPT could track the set of tokens used in response to each query; in practice, that type of logging would become unwieldy and expensive in short order[35]). This, then, is a stronger version of the second possibility outlined above, and ChatGPT would enjoy Section 230 protection for its decision to publish pre-existing data created by another ICP.[36] If the

---

[33] *See* Derek E. Bambauer, *Ghost in the Network*, 162 U. PA. L. REV. 1011, 1019 (2014). Section 230's approach is analogous. Without the statute's safe harbor, platforms would face the untenable choice between potentially ruinous liability for third-party content and costly curation measures that would still be vulnerable to errors creating liability. As with software, policymakers decided that the enormous social benefits of making user-generated content more readily available outweighed the error costs that would inevitably occur.

[34] *See, e.g.*, O'Kroley v. Fastcase, Inc., 831 F.3d 352 (6th Cir. 2016); Batzel v. Smith, 333 F.3d 1018, 1035 (9th Cir. 2003); Maughan v. Google Tech., Inc., 49 Cal. Rptr. 3d 861 (Cal. Ct. App. 2006).

[35] *See* Lanier, *supra* note 6 (discussing potential to track provenance and provide "digital dignity," but without exploring the technological mechanisms and challenges of such an approach).

[36] Since ChatGPT was trained solely on material available via the Internet, the possibility of liability for publishing offline material does not arise. *See* Batzel v. Smith, 333 F.3d 1018, 1032–35

defamatory statement at issue is present in ChatGPT's corpus, and it appears in identical form (or very nearly so), courts seem likely to treat the bot as having retrieved the answer rather than creating it, whatever the underlying mechanics of the code may be. This may also create potentially useful incentives for ChatGPT's developers to curate responses: If they can find underlying defamatory material in the training data, perhaps in response to accusations of defamation, they can both protect themselves using Section 230's shield and also remove that content to prevent future harm.

Fourth, ChatGPT's reply might be partly or completely dependent upon the user's initial query. Imagine a scenario where a neutral question elicits a non-defamatory response, but a leading question causes the bot to produce a defamatory one. The outcome for Section 230 purposes is not immediately clear; there do not appear to be any cases with similar fact patterns. One possible result is that a court would find that ChatGPT was at least partially responsible for the development of the defamatory response. To be treated as a creator or developer, though, ChatGPT would need to make a material contribution to the content—in the Ninth Circuit's framing, it would have to "directly participate in developing the alleged illegality."[37] The other federal circuit courts of appeals have come to similar conclusions via different paths.[38] For example, the Tenth Circuit has held that "a service provider is 'responsible' for the development of offensive content only if it in some way specifically encourages development of what is offensive about the content."[39] To be responsible, the court held, "one must be more than a neutral conduit for that content"; "one is not 'responsible' for the development of offensive content if one's

(9th Cir. 2003). And there are outlier cases that effectively impose publisher-style liability on theories such as inducement. *See, e.g.*, NPS LLC v. StubHub, Inc., No. 06-4874-BLS1, 2009 WL 995483 (Mass. Super. Ct. Jan. 26, 2009).

[37] Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157, 1167–68, 1174–77 (9th Cir. 2008).

[38] *See, e.g.*, Marshall's Locksmith Serv. Inc. v. Google, LLC, 925 F.3d 1263, 1269–71 (D.C. Cir. 2019); Jones v. Dirty World Ent. Recordings, 755 F.3d 398, 410–16 (6th Cir. 2014); Nemet Chevrolet, Ltd. v. Consumeraffairs.com, Inc., 591 F.3d 250, 257–58 (4th Cir. 2009). The Fourth Circuit's latest foray into Section 230 analysis appears to confirm the material contribution approach, although the opinion is otherwise baffling at best. *See* Henderson v. Source for Pub. Data, L.P., 53 F.4th 110, 127–29 (4th Cir. 2022).

[39] FTC v. Accusearch, Inc., 570 F.3d 1187, 1199 (10th Cir. 2009).

conduct was neutral with respect to the offensiveness of the content."[40] ChatGPT is not designed or intended to produce unlawful material such as defamatory statements. Rather, its developers know that the potential exists for the bot to return responses that are, empirically, not true, and that may be damaging to reputation.

The second possibility, foreshadowed slightly in the discussion of responsibility for developing content above, is that courts might treat ChatGPT as a "neutral tool" and thus protected by Section 230.[41] This approach is a classic example of regulating dual-use devices or services: When the tool does not appear specifically designed to produce unlawful results, liability attaches solely to its user.[42] This possibility likely depends upon two factors, one empirical and one psychological. The empirical question is whether ChatGPT is largely a neutral tool, producing defamatory content only when a user guides it in that direction, or whether it generates a cognizable amount of libelous material in response to neutral queries. The psychological question is whether courts see ChatGPT as a new tool, but one in the vein of "standard elements of web sites," or whether its novelty and seemingly customized interactivity lead courts to discard the neutral tools approach for the bot.[43]

This discussion also raises two important collateral possibilities of ChatGPT becoming embroiled in litigation over Section 230. The first is that courts will have to pay closer attention to the role and relative contribution of the user to the generation of defamatory material. In cases such as *Roommates.com*, allocating responsibility was less important because users of the site were expressing restrictions on housing options that might have been unlawful (although both the site and its users ultimately escaped liability based upon an interpretation of the federal Fair Housing Act that excluded renting rooms rather than dwellings). Roommates.com itself encouraged those expressions because it guided users to answer a series of questions that included potentially unlawful answers. ChatGPT, though, is highly context-dependent; it could provide unlawful responses to some queries on a topic but not

---

[40] *Id.*

[41] *See Roommates.com*, 521 F.3d at 1169; Universal Comm. Sys. v. Lycos, Inc., 478 F.3d 413, 419–21 (1st Cir. 2007).

[42] *See Lycos*, 478 F.3d at 421 (describing "standard elements of web sites 'with [both] lawful and unlawful potential'" (internal citation omitted)); *see generally* Sony Corp. of Am. v. Universal City Studios, 464 U.S. 417 (1984).

[43] *Lycos*, 478 F.3d at 421.

others. That gradation will require careful fact-finding and apportionment of liability by courts.

The second consequence is that courts may diverge on how they apply Section 230 to ChatGPT and other generative algorithmic systems. This is already the case, such as with the split in the federal circuit courts of appeals over how to define "intellectual property" as used in Section 230.[44] If these generative tools become widely used, though, consequential differences between courts based upon geography risk undermining one of the statute's core goals, which is to promote the development of this type of interactive service.[45]

Whether Section 230 protects ChatGPT depends importantly on the functionality of the software, the content in its training corpus, and the content of queries posed by users. Section 230 is likely to shield the bot's developers and distributors from a significant share of claims, but assessing when and why immunity applies will become more difficult.

Conclusion

ChatGPT's ability to claim protection under Section 230 varies significantly with how the software performs, now and in the future, and with the content of the material upon which it was trained. If the bot's response, and particularly whether that response is lawful, is strongly dependent upon the query it receives, then the user who issues a query that generates defamatory results may well face liability.[46] Apportioning the relative responsibilities of user and software maker or distributor is likely to be substantively and procedurally complex: Courts will have to learn

---

[44] *Compare id.* at 422–23 (holding that claims for infringement of state-based intellectual property rights are not subject to Section 230 immunity), *with* Perfect 10, Inc. v. CCBill LLC, 488 F.3d 1102, 1119 (9th Cir. 2007) (construing "the term 'intellectual property' to mean 'federal intellectual property'" and thus finding immunity under Section 230 from state trademark claims).

[45] *See* 47 U.S.C. § 230(b)(1).

[46] The user would be an information content provider under Section 230 and thus not immunized under 230(c)(1). Defamation also requires two additional elements. First, there must be publication to a third party, which could occur if the querying user redistributed the response or even if someone else was looking over their shoulder as it was generated. Second, the querying user must possess the requisite state of mind. Although the issue is not fully resolved, Supreme Court precedent suggests that this state of mind must be at least negligence regarding the truth of the statement for private figures, and for public figures, the famous actual malice standard applies. *See* Dun & Bradstreet, Inc. v. Greenmoss Builders, Inc., 472 U.S. 749 (1985); New York Times v. Sullivan, 376 U.S. 254 (1964).

something about large language models and pre-trained transformers, and plaintiff-side counsel may have to join individual defendants to effectively pursue developers or platforms.

Section 230 has considerably more potential as a source of immunity for ChatGPT than most commentators appreciate. However, the reasons for the conventional wisdom are also a source of risk for the bot: Courts, too, may see the software as quasi-sentient and perhaps even malignant, given its penchant for generating false, random, and at times incomprehensible answers. This impression is furthered by the heated if not somewhat hysterical coverage of ChatGPT in the media, against the backdrop of dystopian science fiction that illustrates the possible, albeit unlikely, risks of powerful artificial intelligence.

As a policy matter, Section 230 ought to cover ChatGPT, particularly in its early development. A key purpose of the statute was and remains to encourage the development of interactive Internet technologies without fear of ruinous legal liability. And the actual risk of harm from defamatory statements (which is a separate question from immunity, of course) seems quite low. Few take ChatGPT responses seriously at present, given its reputation as a hallucination engine, and that number falls further among observers who understand something about machine learning. OpenAI and the developers of similar generative tools ought to continue to emphasize that their software is utterly dependent upon its training data and is explicitly not designed for accuracy, let alone truth. ChatGPT's potential as an interactive tool looks promising, even if only to generate first drafts, and the potential harms from it are greatly overblown, at least at present. In short, Section 230 has significant capacity to shield ChatGPT from liability for defamation, and that protection seems socially desirable.