2024

Risky Speech Systems: Tort Liability for AI-Generated Illegal Speech

Margot E. Kaminski

Risky Speech Systems: Tort Liability for AI-Generated Illegal Speech

Author : Margot Kaminski

Date : February 6, 2024

- Jane Bambauer, *Negligent AI Speech: Some Thoughts about Duty*, 3 **J. Free Speech L.** 344 (2023).
- Nina Brown, *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*, 3 **J. Free Speech L.** 389 (2023).

How should we think about liability when AI systems generate illegal speech? The *Journal of Free Speech Law*, a peer-edited journal, ran a topical 2023 symposium on Artificial Intelligence and Speech that is a must-read. This JOT addresses two symposium pieces that take particularly interesting and interlocking approaches to the question of liability for AI-generated content: Jane Bambauer's *Negligent AI Speech: Some Thoughts about Duty*, and Nina Brown's *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*. These articles evidence how the law constructs technology: the diverse tools in the legal sensemaking toolkit that are important to pull out every time somebody shouts "disruption!"

Each author offers a cogent discussion of possible legal frameworks for liability, moving beyond debates about First Amendment coverage of AI speech to imagine how substantive tort law will work. While these are not strictly speaking First Amendment pieces, exploring the application of liability rules for AI is important, even crucial, for understanding how courts might shape First Amendment law. First Amendment doctrine often hinges on the laws to which it is applied. By focusing on substantive tort law, Bambauer and Brown take the as-yet largely abstract First Amendment conversation to a much-welcomed pragmatic yet creative place.

What makes these two articles stand out is that they each address AI-generated speech that is illegal—that is, speech that is or should be unprotected by the First Amendment, even if First Amendment coverage extends to AI-generated content. Bambauer talks about speech that physically hurts people, a category around which courts have been conducting free-speech line-drawing for decades; Brown talks about defamation, which is a historically unprotected category of speech. While a number of scholars have discussed whether the First Amendment covers AI-generated speech, until this symposium there was little discussion of how the doctrine might adapt to handle liability for content that's clearly unprotected.

Bambauer's and Brown's articles are neatly complimentary. Bambauer addresses duties of care that might arise when AI misrepresentations result in physical harm to a user or third parties. Brown addresses a products-liability approach to AI-generated defamation. Another related symposium piece that squarely takes on the question of liability for illegal speech is Eugene Volokh's *Large Libel Models? Liability for AI Output*. The Brown and Bambauer pieces speak more directly to each other in imagining and applying two overlapping foundational liability frameworks, while Volokh's piece focuses on developing a sui generis version of developer liability called "notice-and-blocking" that he grounds in Brown's idea of using products liability as a starting point. That is, Bambauer and Brown provide the necessary building blocks; Volokh's article is an example of how one might further manipulate them.

Bambauer writes of state tort liability, as it might be modified by state courts incorporating free speech

values. She explains that she has "little doubt that the output of AI speech programs will be covered by free speech protections" (P. 347) (as do my co-authors and I) but also that "the First Amendment does not create anything like an absolute immunity to regulatory intervention," especially when it comes to negligence claims for physical harm. (P. 348.) Bambauer convincingly claims that the duty element of negligence is where the rubber will hit the road in state courts when it comes to determining the right balance between preventing physical harms and protecting free speech values. She identifies different categories of duty as an effective way of categorizing existing cases that address analogous problems (from books that mis-identify poisonous mushrooms as edible, to doctors who provide dangerously incorrect information to patients).

Bambauer divides her discussion of duty into three broad categories, followed by additional subcategories: 1) situations where AI systems provide information to a user that causes physical harm to that user; 2) situations where AI systems provide information to a user who then causes physical harm to a third party; and 3) situations where AI systems would have provided accurate information that could have averted harm, had a user consulted them (reminiscent of Ian Kerr's and Michael Froomkin's prescient work on the impact of machine learning on physician liability). Throughout, this article is logical, clearly organized, factually grounded, and neatly coherent, even where a reader might depart from its substantive claims.

These categories allow Bambauer to tour the reader through available analogies, comparing AI "to pure speech products, to strangers, or to professional advisors" and more. (P. 360.) If an AI system's erroneous output is analogized to a book, Bambauer argues that developers will not and should not be found liable, as with a book that misidentified poisonous mushrooms as edible as in the Ninth Circuit's *Winter* case. (Eerily, this exact fact pattern has already arisen with AI-generated foraging books.) If, under different factual circumstances, AI-generated content is more appropriately analogized to professional advice in a specialized domain such as law or medicine, there might be a higher duty of care. Or, courts might use a "useful, if strained analogy" of "wandering precocious children," where parents/developers might be held liable under theories of "negligent supervision" for failing to anticipate where their child/generative AI might be doing dangerous things. (P. 356.) This might, Bambauer muses, nudge courts to focus on what mechanisms an AI developer has put in place to find and mitigate recurring harms. This is a classic "which existing analogy should apply to new things?" article, but done well. Others might take this logic further by pulling analogies from other spaces (I'm thinking here for example of Bryan Choi's work on car crashes, code crashes, and programmers' duties of care).

This takes us to Brown's intervention. Brown examines defamation claims through a products liability lens, asking what interventions a developer might be required to take to mitigate the known risk of defamatory content. Brown starts with a summary of how chatbots work, so the rest of us don't have to. (I will be citing this section often.) She quickly and clearly explains the defamation puzzle: that the current law focuses largely on the intent of the speaker/publisher of defamatory content. This approach runs into issues when we are talking about the developers of AI systems, who Brown argues will almost never have the requisite intent under current defamation law.

Brown then turns to dismantling hurdles to a products liability approach (is it a product? What's the role of economic loss doctrine?). Readers may find this part more or less convincing, but resolving the hurdles (it's a product, she thinks economic loss doctrine is not a problem) allows her to get to the really interesting part of the article: what substantive duties a developer might have, if AI-generated defamation gets framed as a products liability problem. Brown argues that "a design defect could exist if the model was designed in a way that made it likely to generate defamatory statements." (P. 410.) She provides concrete examples grounded in current developer practices: the use of flawed datasets rife with false content; the prioritization of sensational content over accuracy; a failure to take steps to

reduce the likelihood of hallucinations; a failure to test the system.

I'm still not sure a products liability approach will survive the Supreme Court's recent emphasis on scienter in First Amendment cases, but one can hope. In several recent cases, most prominently in Counterman v. Colorado, the Supreme Court has insisted on a heightened intent standard for unprotected speech in order to protect speakers from a chilling effect that occurs if one cannot clearly determine whether one's speech is unprotected or protected.[1] In *Counterman*, the unprotected speech category at issue was true threats, which the Court found could not be determined under an objective standard but required a query of speaker intent. The Court reasoned that a heightened intent standard creates a penumbra of protection for borderline speech that is close to but not unprotected speech—such as opinionated criticism of a public figure bordering on defamation, or vigorous political speech at a rally bordering on incitement. Brown presents the products-liability approach as a sort of hack to get around the specific intent requirement of "actual malice" for defamation of public figures (private figures require only negligence, but arguably a heightened form of it). She does not really inquire about whether this is possible—whether today's Court, post-*Counterman*, would accept this move. I personally think there is space in the Court's reasoning in *Counterman* for moving away from specific intent, but it would have been nice to know Brown's thoughts.

Together, these two articles offer a trio of important contributions: foundations for First Amendment debates about unprotected speech and AI systems; creative but grounded ways of imagining duties of care in the context of developer liability (relevant, too, to evolving discussions of platform liability); and an important basis for discussions about the role of tort law in establishing risk mitigation for content-generating AI systems in the U.S. legal context. Regulators have increasingly defaulted to a regulatory approach to risk mitigation for AI systems, including or especially in the EU. If, as is likely, the United States fails to enact its counterpart to, the Digital Services Act (DSA), Europe's massive new law regulating content moderation, tort law may be where AI risk mitigation plays out in the United States.

1. *Counterman v. Colorado*, 600 U.S. 66 (2023).

Cite as: Margot Kaminski, *Risky Speech Systems: Tort Liability for AI-Generated Illegal Speech*, JOTWELL (February 8, 2024) (reviewing Jane Bambauer, *Negligent AI Speech: Some Thoughts about Duty*, 3 **J. Free Speech L.** 344 (2023); Nina Brown, *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*, 3 **J. Free Speech L.** 389 (2023)), https://cyber.jotwell.com/risky-speech-sys…d-illegal-speech/.