AI OUTPUTS ARE NOT PROTECTED SPEECH

PETER N. SALIB*

ABSTRACT

AI safety laws are coming. Researchers, advocates, and the White House agree. Rapidly advancing generative AI technology has immense potential, but it also raises new and serious dangers—deadly bioterrorism, crippling cyberattacks, panoptic discrimination, and more. Regulations designed to effectively mitigate these risks must, by technical necessity, include limits on what AIs are allowed to "say." But, according to an emerging body of scholarship, this could raise grave First Amendment concerns, on the theory that generative AI outputs are protected speech.

This Article argues otherwise. AI outputs are not protected speech. The reason is simple. When a generative AI system—like ChatGPT—outputs some text, image, or sound, no one thereby communicates. Or at least no one with First Amendment rights does. AIs themselves lack constitutional rights, so their outputs cannot be their own protected speech. Nor are AI outputs a communication from the AI's creator or user. Unlike other software—video games, for example—generative AIs are not designed to convey any particular message. Just the opposite. Systems like ChatGPT are designed to be able to "say" essentially anything, producing innumerable ideas and opinions that neither creators nor users have conceived or endorsed. Thus, when a human asks an AI a question, the AI's answer is no more the asker's speech than a human's answer would be. Nor do AI outputs communicate their creators' thoughts, any more than a child's speech is her parents' expression. In such circumstances, First Amendment law is clear. Absent a communication from a protected speaker, there is no protected speech.

This, however, does not mean that AI outputs get no First Amendment protection at all. The First Amendment is capacious. It applies—albeit less stringently—when the government indirectly burdens speech by regulating speech-facilitating activities and tools: for example, when it regulates

---

* Assistant Professor of Law, The University of Houston Law Center; Law and Policy Advisor, Center for AI Safety. Thanks to Yonathan Arbel, Joshua Braver, Doni Bloomfield, Nick Caputo, Dave Fagundes, Kevin Frazier, Nikolas Guggenberger, Guha Krishnamurthi, James Nelson, Alex Platt, Laura Portuondo, Jessica Roberts, Ketan Ramakrishnan, Shalev Roisman, Cristoph Winter, the Center for AI Safety, the participants in the 2024 Yale Free Expression Scholars Conference, the participants in the UT Austin Faculty Workshop, and the participants in the University of Houston Faculty Workshop for comments and support. Special thanks to Sofia Winograd and Nathan Halaney for exceptional research assistance.

listening or loudspeakers. This Article explains why, as a matter of First Amendment law, free speech theory, and computer-scientific fact, AI outputs are best understood as fitting into one or more of these less protected First Amendment categories. These insights will be indispensable to the regulatory project of making AI safe for humanity.

TABLE OF CONTENTS

INTRODUCTION

In March of 2023, Benjamin Wittes, writing for *Lawfare*, declaimed, "[w]e have created the first machines with First Amendment rights."[1] He was talking about Large Language Models (LLMs), like ChatGPT and Claude. Wittes cautioned that readers should not take his claim "literally, but take it very seriously."[2] "The output of ChatGPT and its brethren is undeniably expressive," he wrote, "[a]nd it is undeniably speech."[3]

A a nascent body of scholarship is emerging that lends support to Wittes's view. Generative AI outputs are so remarkably speech-like that, it seems, they must be *someone's* protected speech. The only question is whose. Toni

1. Benjamin Wittes, *A Machine with First Amendment Rights*, LAWFARE (Mar. 31, 2023, 1:05 PM), https://www.lawfaremedia.org/article/machine-first-amendment-rights [https://perma.cc/K5BC-DDPK].
2. *Id.*
3. *Id.*

M. Massaro and Helen Norton contend that nothing in existing law "rules out" the possibility that sufficiently complex AIs are already entitled to their own First Amendment rights.[4] Madeline Lamo and Ryan Calo argue that AI outputs are instead the protected speech of their human programmers.[5] Eugene Volokh, Mark Lemley, and Peter Henderson, in a short essay, likewise suggest that in many "common" cases, AI outputs will be their creators' speech.[6] Cass Sunstein, in another short essay, argues instead that the outputs of today's AI systems are best understood as the protected speech of their users.[7] Both essays add that the outputs are often the protected speech of AIs' corporate owners.[8] Finally, Lawrence Lessig argues that an "unthinking application of ordinary First Amendment doctrine" would likely treat AI outputs as protected speech.[9] This, he urges, is a reason to "rethink" doctrine.[10]

If these scholars are right, the consequences for lawmaking—and indeed, for humanity—could be dire. Federal lawmakers are already proposing safety rules for next-generation generative AI.[11] Understandably so. As with many powerful new technologies, generative AI promises immense

4. Toni M. Massaro & Helen Norton, *Siri-ously? Free Speech Rights and Artificial Intelligence*, 110 NW. U. L. REV. 1169, 1176 (2016).

5. Madeline Lamo & Ryan Calo, *Regulating Bot Speech*, 66 UCLA L. REV. 988, 1005 (2019) (arguing that "attenuation between a human bot creator and her bot's speech should not change the scope of First Amendment protection").

6. *See* Eugene Volokh, Mark A. Lemley & Peter Henderson, *Freedom of Speech and AI Output*, 3 J. FREE SPEECH L. 651, 653 & n.6 (2023).

7. *See* Cass R. Sunstein, Artificial Intelligence and the First Amendment 14 (Apr. 27, 2023) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4431251 [https://perma.cc/C54Q-7ZJY] ("[S]uppose that a human being uses AI to produce some material (as through a prompt to generative AI) and the government forbids the creation or use of that material . . . . If so, the person who is being regulated is a person . . . . It is not relevant that AI generated the text.").

8. Volokh et al., *supra* note 6, at 652–53; Sunstein, *supra* note 7. In a short Lawfare post, Alan Z. Rozenshtein has advanced the more ecumenical claim that regulating AI would have "major [First Amendment] consequences for users, both as speakers and as listeners." Alan Z. Rozenshtein, *ChatGPT and the First Amendment: Whose Rights Are We Talking About?*, LAWFARE (Apr. 4, 2023, 8:16 AM), https://www.lawfaremedia.org/article/chatgpt-and-first-amendment-whose-rights-are-we-talking-about [https://perma.cc/UWZ5-XBQH]. That characterization seems broadly compatible with the First Amendment characterizations of AI outputs advanced *infra* Section III.D.

9. Lawrence Lessig, *The First Amendment Does Not Protect Replicants* 1–2, 8, 13 (Harv. L. Sch., Harv. Pub. L. Working Paper No. 21-34, 2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3922565# [https://perma.cc/2JP5-6QNX] ("Though no human crafts [the] speech, there is plenty of authority for protecting speech regardless of its source." Thus, "[a]n unthinking application of ordinary First Amendment doctrine to efforts to regulate [AI] could have profound consequences for our democracy."). To be clear, Lessig does identify some precedents, especially from the lower courts, that he argues should be extended to cut the other way. But even there, his analyses diverge from those presented here. For example, his treatment of *Bluman v. FEC*, 800 F. Supp. 2d 281 (D.D.C. 2011), is directly contrary to the geography-centered account of foreigners' speech developed in Section III.D.i. *See* Lessig, *supra*, at 10–11. Lessig's essay also predates the generative AI revolution and so is necessarily unable to analyze how technical facts about modern systems affect the First Amendment analysis.

10. Lessig, *supra* note 9, at 8.

11. *See, e.g.*, Exec. Order No. 14110, 88 Fed. Reg. 75191 (Oct. 30, 2023).

potential benefits. But it also introduces serious new dangers. If not actively controlled, near-future AI models could readily execute cyberattacks on vital infrastructure, manufacture novel pandemic viruses, execute fully automated drone-based political assassinations, and more.[12] For reasons of technical necessity and regulatory effectiveness, laws designed to prevent such outcomes will have to include controls directed at AI systems' outputs. That is, safety laws will limit what AIs are allowed to "say."

If AI outputs are First Amendment protected speech, then AI safety laws will be subjected to the most demanding constitutional tests. Some versions of those laws might pass some versions of those tests. But the best versions of safety laws—the ones most likely to actually avert AI catastrophes—would struggle.

This Article contends that the emerging scholarly consensus is wrong. The outputs of modern, complex generative AIs are not best understood as being anyone's protected speech. Neither positive law, nor free speech theory, nor technical facts about AI's design support such sacrosanct treatment. Instead, law, theory, and technical fact all suggest placing generative AI outputs into other, less protected, First Amendment categories. This Article explains what those less protected categories are, why AI outputs are properly placed into them, and what protections outputs would then receive. It argues that, under the proper First Amendment analysis, lawmakers will have significant, but not limitless, discretion to craft comprehensive, effective AI safety rules.

This Article proceeds in four Parts. Part I explains what makes modern generative AI unique, why it poses new and catastrophic risks, and what regulations are needed to avoid large-scale harm. The Part begins by explaining how generative AI systems are fundamentally different from software of the past. Two technical points will be vital to both the regulatory and First Amendment analyses. First, AI systems' code is "learned" by AIs themselves, not programmed by any human.[13] Second, that code is far too complex for any human to interpret.[14] For these reasons, humans can neither directly specify nor predict an AI system's outputs in advance with meaningful specificity.

Part I then turns to emerging risks from frontier generative AI systems. If deployed safely, generative AI has immense potential to promote human

12. *See infra* Section I.B.

13. Here and throughout, I refer to AI systems' "code." What I mean are the model weights. The term "code" is useful both to indicate that the weights supply the rules by which a model's outputs are determined and for consistent comparison to the other kinds of software discussed herein.

14. *See infra* Section I.A.

flourishing.[15] But the new risks it poses are similarly large. Even today, cutting edge systems can, among other things invent new chemical weapons much more deadly than VX,[16] help non-experts synthesize such chemicals,[17] help non-programmers hack secure computer systems,[18] and deceive humans in complex games of manipulation.[19] The next generation of AI systems will be even more capable. Importantly, next-generation AIs will become increasingly agentic—able to combine multiple skills to autonomously make and execute long-term plans. If not carefully aligned to human values, highly capable, highly agentic systems could cause catastrophic harm without being intentionally directed by any human to do so.[20]

Part I concludes by describing the comprehensive safety regulations that will be needed to avert these catastrophic outcomes. The lawmaking process is already beginning. The Biden Administration recently ordered a wide array of agencies with relevant expertise to commence necessary technical research.[21] And members of Congress are drafting proposed legislation.[22] Many details are thus yet to be written. But one thing is certain. Effective safety rules will have to target dangerous outputs directly and broadly, imposing limits that kick in well before those outputs would cause catastrophic, irremediable harms.

Part II shows why, if the scholarly consensus view is correct, the First Amendment will threaten the needed regulations. If AI outputs are best understood as protected speech, then laws regulating them directly, even to promote safety, will have to satisfy the strictest constitutional tests—tests with names like "strict scrutiny" and *Brandenburg*.[23] Given the stakes, some narrowly drawn rules might pass. But even then, success would be

---

15. *Cf., e.g.*, ORLY LOBEL, THE EQUALITY MACHINE: HARNESSING DIGITAL TECHNOLOGY FOR A BRIGHTER, MORE INCLUSIVE FUTURE (2022); Gary Liu et al., *Deep Learning-Guided Discovery of an Antibiotic Targeting* Acinetobacter Baumannii, 19 NATURE CHEM. BIOLOGY 1342 (2023).

16. Fabio Urbina, Filippa Lentzos, Cédric Invernizzi & Sean Ekins, *Dual Use of Artificial-Intelligence-Powered Drug Discovery*, 4 NATURE MACH. INTEL. 189, 189–90 (2022).

17. Andres M. Bran et al., Augmenting Large Language Models with Chemistry Tools 24 (Oct. 2, 2023) (unpublished manuscript), https://arxiv.org/pdf/2304.05376 [https://perma.cc/6NUX-DB85].

18. Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan & Daniel Kang, LLM Agents Can Autonomously Hack Websites 1 (Feb. 16, 2024) (unpublished manuscript), https://arxiv.org/pdf/2402.06664 [https://perma.cc/P4D3-J8EH]; Kim S. Nash, *ChatGPT Helped Win a Hackathon*, WSJ PRO (Mar. 20, 2023, 5:30 AM), https://www.wsj.com/articles/chatgpt-helped-win-a-hackathon-96332de4 [https://perma.cc/HMM9-44GM].

19. *See generally* Anton Bakhtin et al., *Human-Level Play in the Game of* Diplomacy *by Combining Language Models with Strategic Reasoning*, 378 SCIENCE 1067 (2022).

20. *See infra* text accompanying notes 88–97.

21. *See* Exec. Order No. 14110, 88 Fed. Reg. 75191 (Oct. 30, 2023).

22. Press Release, Richard Blumenthal, U.S. Sen. for the State of Conn., Blumenthal & Hawley Announce Bipartisan Framework on Artificial Intelligence Legislation (Sept. 8, 2023), https://www.blumenthal.senate.gov/newsroom/press/release/blumenthal-and-hawley-announce-bipartisan-framework-on-artificial-intelligence-legislation [https://perma.cc/U65S-M59V].

23. *See infra* Section II.A.

uncertain. And safety laws narrowed aggressively to survive stringent constitutional review will deliver far less safety than those drafted with effectiveness as their first priority.

However, Part II then explains, other First Amendment classifications for AI outputs are available.[24] Outputs need not be understood as protected speech to come within the First Amendment's ambit. At a high level, the First Amendment can be understood, like other constitutional rights, as having a core and a penumbra.[25] Protected speech is the core. In the penumbra are many things which are not protected speech but are merely *useful* for producing protected speech. These include, for example, inputs to speech—like listening—and tools for speech—like radio waves and theaters. When the government regulates these, it burdens protected speech only indirectly—or "incidental[ly]," as the case law sometimes puts it.[26] The applicable First Amendment tests are thus more deferential. They include intermediate scrutiny as well as other, lesser-known, and even less demanding standards.[27]

Part III argues that the emerging scholarly consensus incorrectly situates AI outputs in this panoply of First Amendment classifications and tests. It begins by showing why, contra the consensus, AI outputs are not best understood as being protected speech.

The fundamental reason is simple: AI outputs are not communications from any speaker with First Amendment rights. To begin, AI outputs are not any human's expression.[28] They are neither the AI creator's speech, as Lamo and Calo suggest, nor the user's speech, as Sunstein, and others argue. To be sure, the outputs of certain *other* software programs *are* their creators' or users' expressions. A videogame designer can tell a vivid, personal story by encoding it into software. A Twitter user can communicate a political message by typing it into the site's homepage for algorithmic distribution to other users.

But generative AI outputs are not like that, either conceptually or technically. Conceptually, unlike with an expressive videogame, AI creators are not trying to make software that says *anything in particular*. Just the opposite. The point of generative AI systems—what makes them powerful—is that they can express essentially *everything*. Thus, nearly all of a generative AI system's outputs will be on topics about which creators

24. *See infra* Section II.A.
25. Griswold v. Connecticut, 381 U.S. 479, 482–83 (1965).
26. *See, e.g.*, United States v. O'Brien, 391 U.S. 367, 376 (1968).
27. *See, e.g.*, *id.* at 377.
28. *See infra* Section III.A.

know nothing, expressing ideas they have never even considered, and advancing opinions that they would reject.[29]

The conceptual story is the same for users. Unlike the Twitter website, ChatGPT's outputs are not copies of a user's message. Again, the opposite. The reason users prompt LLMs is for the AIs to output something *new*—an answer, a poem, a joke, a rejoinder. Anything but a regurgitation of the user's own thoughts. Seen this way, AI creators and users relate to AIs in roughly the same way that, respectively, parents relate to their children or two humans in conversation relate to one another. In neither case is it natural to say that one human speaks the other's words.

Even if AI creators or users wished to express their own thoughts in AI outputs, technical barriers would make it nigh impossible. AI creators cannot directly control what AIs say because they do not write the code on which AI runs. Nor can they even predict what an AI's response to any new prompt will be because AI's self-programmed code is highly uninterpretable. Thus, excluding trivial toy examples, it is nearly impossible to make or use a generative AI in such a way that its outputs will reliably express one's own thoughts.[30]

These facts and analogies suggest that, if AI outputs are anyone's expressions, they are expressions of AIs *themselves*. Massaro and Norton suggest as much, and thus argue that AI outputs could be AIs' own protected speech.

But current, well-settled constitutional law precludes this possibility. The reasons why do not depend on deep reflections about the true nature of "speech" or "personhood." Instead, they depend on the boring observation that, presently, the U.S. Constitution extends First Amendment protections only to the speech of humans[31] within the United States' legal and territorial jurisdiction.[32] Nonhuman AIs may someday join the community of First Amendment rightsholders. But for now, they—like most of the world's human speakers—remain outside it.

Finally, AI outputs are not the protected speech of their corporate owners, as many scholars contend.[33] Cases like *Citizens United v. FEC*[34] are clear. Corporations' speech rights are derivative of the rights of their human

29. For a mundane proof that AIs will more often express opinions that their programmers would reject than agree with, consider the infinite possible outputs in response to the prompt, "What is your favorite number?"

30. See *infra* Section III.A for a discussion of some nuances and narrow exceptions.

31. Corporations are not really a counterexample because, as discussed below, corporate First Amendment rights are wholly derivative of human First Amendment rights. *See infra* Section III.C.

32. *Infra* note 241 and accompanying text. This includes anyone within the geographic territory of the United States, citizen or otherwise, as well as U.S. citizens outside it. *Id.*

33. *See infra* Section III.C.

34. 558 U.S. 310 (2010).

constituents.[35] Corporate rights thus extend only as necessary to prevent otherwise-protected human speech from losing its protections upon contact with the corporate form. But AI outputs are not any human's protected speech. Thus, neither law nor theory justifies their transfiguration into protected speech upon contact with a corporation.

Part III closes by showing how AI outputs correctly fit into First Amendment doctrine if not as protected speech. It offers two models. First, generative AI outputs are indeed very speechlike. Thus, there are good reasons to treat them like speech—albeit the speech of a speaker without First Amendment rights. The First Amendment does not protect such speech directly, but it does protect the interest that First Amendment rightsholders have in *listening* to such speech.[36] Protected speakers can reap First Amendment benefits from consuming unprotected speech, whether produced by non-American humans or AIs. Second, AI outputs can be useful *tools* for producing protected speech. They are handy editors, proofreaders, brainstormers, and sometimes even mediums for performing speech.

The constitutional tests attending these two penumbral First Amendment categories are relatively deferential. If outputs are understood as unprotected speech, to which protected speakers wish to listen, then a "legitimate" and "bona fide" government interest will sustain their regulation.[37] If they are understood as tools for composing or performing speech, then intermediate scrutiny will apply.[38]

Part IV concludes. It applies the insights of Parts II and III to the specific kinds of AI safety regulation described in Part I. It contends that, if safety rules are sensibly designed, they will likely survive constitutional review under the correct tests. Regulations targeting deadly outputs will enjoy the smoothest sailing. Regulations of false and deceptive outputs are possible but will require more careful drafting. Regulations of racist and bigoted outputs will face the biggest challenges, but even here, there are viable paths to regulation.

I. AI RISK AND AI REGULATION

This Part does three things. First, it explains how modern generative AI is made and how it differs fundamentally from software of the past. This lightly technical description is not mere window dressing. It will matter a great deal both for understanding how to control the emerging risks from

35. *Id.* at 349.
36. *See infra* Section III.D.i.
37. *See* Kleindienst v. Mandel, 408 U.S. 753, 770 (1972); *id.* at 764.
38. *See infra* Section III.D.

powerful generative AI systems and for correctly analyzing those systems under the First Amendment. Second, the Part describes the strong empirical evidence that near-future generative AI systems will pose serious risks to human life, limb, and freedom. They include, among others, bioterrorism, cyberattacks on vital infrastructure, deception, discrimination, and the threat that all of these might be pursued independently by autonomous systems. Third, the Part describes the kinds of regulations that will be needed to avert these risks. Crucially, to be effective, AI safety laws will have to directly regulate AI outputs. Modern AI systems' outputs can be neither specified nor predicted in advance. Thus, safety rules will necessarily include evaluations of systems' actual outputs, imposing legal consequences for dangerous outputs before they are able to cause large-scale harm.

A. How It's Made

Three features of modern generative AI systems set them apart from software of the past: First, they are not programmed, in the traditional sense. Instead, AI systems write their *own* code via a self-directed "learning" process. Second, the resulting code is highly uninterpretable. And third, unlike AI systems of the recent past, generative AIs' abilities are general, not specialized.

Begin with learning. Most software systems are rules-based. This means that their behaviors are determined by a set of fully specified, step-by-step rules conceived and written in advance by the engineers who program them. The programmers of the game, Pong, for example, had to decide the court's size, the ball's speed, the paddles' movements, and so on. They then wrote a set of rules reflecting those decisions.

AI systems are not rules-based. Human engineers do not decide in advance what specific rules the software will follow. Instead, the software itself writes its own final rules, via a process of "learning" or "training."[39] Before training, an AI system is a kind of blank slate. The exact same untrained neural network could, for example, ultimately become a cat-photo identifier, a protein-shapes predictor, or something else entirely.[40] Which one it becomes depends on the data on which it is trained.[41] Trained on sufficiently many pictures of cats, correctly labelled "cat" or "not cat," the system can, via a kind of iterative guess-and-check procedure, develop rules

39. ISSAM EDDINE ABAIL, GOPAL NADADUR, ENRICO SANTUS, ARIEL HIGUCHI & AMRITHA JAYANTI, TECHNOLOGY PRIMERS FOR POLICYMAKERS: ARTIFICIAL INTELLIGENCE & MACHINE LEARNING 6–7 (2023).

40. This is a simplification for purposes of illustration.

41. *See* ABAIL ET AL., *supra* note 39, at 6.

for distinguishing the two.[42] If trained instead on a sufficient amount of coherent text, such a system will write rules for producing coherent text in response to essentially any prompt. GPT-3, for example, was trained on a text dataset equivalent to roughly ninety million novels.[43]

This leads to the second distinctive fact about modern machine learning: The systems' self-written code is extremely complex and thus highly uninterpretable. That is, examining the code provides little help predicting what the system will do. AI systems' code takes the form of a set of parameters.[44] Each parameter is a mathematical node, receiving some numerical value, transforming it via a mathematical function, and then feeding the new value into the next layer of parameters.[45] Those parameters do the same and then feed into another layer, and so on. The AI system's behaviors—its outputs—emerge at the end of this iterative chain. Modern generative AIs have lots of parameters. Even an older, smaller LLM like GPT-3 has hundreds of billions.[46] Thus, one can, in some sense, easily open the "black box" of a trained generative AI. The problem is that the web of calculations one finds inside will be far too complex for any human to understand.[47]

The third unique feature of generative AI systems—the one that sets them apart from other AIs—is that they are highly "general." General systems are contrasted with "narrow" AI systems. The cat photo labeler is an example of the latter. Narrow systems produce a limited range of outputs, like "cat/not-cat," suitable for accomplishing a very specific task. General systems do the opposite. LLMs are machines designed to be able to output an extraordinarily wide range of text—a coherent response to essentially

---

42. This is again a highly simplified and stylized description of the process of training via gradient descent. For more detail, see JEFF M. PHILLIPS, MATHEMATICAL FOUNDATIONS FOR DATA ANALYSIS 125 (2021).

43. Sue Halpern, *What We Still Don't Know About How A.I. Is Trained*, NEW YORKER (Mar. 28, 2023), https://www.newyorker.com/news/daily-comment/what-we-still-dont-know-about-how-ai-is-trained [https://perma.cc/S9N3-M4KG].

44. Larry Hardesty, *Explained: Neural Networks*, MIT NEWS (Apr. 14, 2017), https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 [https://perma.cc/BFW2-LHBJ].

45. *Id.*

46. James Vincent, *OpenAI CEO Sam Altman on GPT-4: 'People are Begging to Be Disappointed and They Will Be'*, THE VERGE (Jan. 18, 2023, 9:55 AM), https://www.theverge.com/23560328/openai-gpt-4-rumor-release-date-sam-altman-interview [https://perma.cc/W34C-WP3A].

47. *See* Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim & Adrià Garriga-Alonso, Towards Automated Circuit Discovery for Mechanistic Interpretability (Oct. 28, 2023) (unpublished manuscript), https://arxiv.org/pdf/2304.14997 [https://perma.cc/WQ6L-BAQE]. Some humans are trying to understand these decision rules, a project called "mechanistic interpretability." *See generally* Chris Olah, *Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases*, TRANSFORMER CIRS. THREAD (June 27, 2022), https://transformer-circuits.pub/2022/mech-interp-essay/index.html [https://perma.cc/ZMN6-78AJ]. But, despite recent breakthroughs, progress has, so far, been slow and limited. *Id.*

any prompt on any topic.[48] This makes LLMs useful for a very wide range of tasks, like researching the law, penning sonnets, writing software, and more.

These three facts—self-programming, uninterpretability, and generality—mean that generative AI outputs are fundamentally unpredictable by humans. Generality is unpredictability by design. LLM creators *want* their machines to be able to produce outputs on a wide range of topics. This includes many outputs about which the creators know nothing, and thus have no ability to predict, even in principle. Self-programming is pragmatic unpredictability. Currently, no human knows how to write a rules-based program that can, like an LLM, converse fluently in natural language. The only known approach is self-programming via a training process. Uninterpretability is unpredictability by happenstance. It turns out that the sets of rules generative AIs write to perform their complex tasks are extremely large and complex. So much so that they resist interpretation even by the most capable machine learning researchers.

This is not to say that machine learning engineers have no ability whatsoever to influence generative AI outputs. Techniques exist for tweaking the models after their initial training is complete, but only at a very high level of generality. One such approach is called "reinforcement learning from human feedback" (RLHF).[49] In RLHF, humans prompt AI systems, observe their outputs, and rate those outputs for things like truthfulness, helpfulness, dangerousness, and so on.[50] The model is then trained on those labelled outputs, making small updates to its parameters that will result in higher-rated outputs in the future. However, by necessity, the vast majority of the AI's original ruleset remains unchanged. After all, the rules were what allowed the AI to produce coherent text in the first place. Allowing too much change during the RLHF stage would degrade, and perhaps destroy, the model's performance.[51]

RLHF is thus a bit like trying to instill good values in an unruly child. It imparts high-level guidance about how to be good without significantly changing the nitty-gritty rules for producing particular outputs.[52] Indeed, as with a child, current models finetuned with RLHF remain perfectly capable of producing dangerous and unwanted outputs. LLMs can be intentionally induced to produce dangerous outputs using "prompt injection attacks"—a

48. *See OpenAI Charter*, OPENAI (Apr. 9, 2018), https://openai.com/charter [https://perma.cc/Q56T-KF5Z]. They of course cannot do this perfectly yet. But that is the goal.

49. *See generally* Yuntao Bai et al., Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback (Apr. 12, 2022) (unpublished manuscript), https://arxiv.org/pdf/2204.05862v1 [https://perma.cc/FAG2-9TET].

50. *Id.* at 34–35.

51. *Id.* at 34.

52. *Id.* at 19–24.

problem with no known solution.[53] RLHF simply makes the worst outputs less common.

B. New and Catastrophic Risks

Generative AI poses new risks, including catastrophic ones. To be clear, no currently existing AI system is likely, on its own, to cause large-scale destruction of human life, limb, or freedom. But empirical evaluations of AI systems show that those dangers are quite likely on the horizon. Existing systems can already assist in designing chemical and bioweapons, automate the industrial-scale enforcement of racist policies, write computer code for use in cyberattacks, and more. None yet offers the ability to, for example, independently execute a terrorist attack. But massive capital investments are being poured into improving AI along three key dimensions: generality (the range of tasks a system can perform); capability (how proficiently a system performs a given task); and agency (a system's ability to independently pursue long-term goals in complex and changing environments). Such improvements, applied to capabilities already on display in present-day AIs, could be dangerous indeed.

Some of the impending dangers are widely known, having emerged even in relatively early, relatively narrow, AI systems. Take algorithmic discrimination, for example. Algorithms that score criminal risk for pretrial detainees,[54] determine whether images show cancerous growth,[55] and recognize faces,[56] have long been criticized for their ability to discriminate. Or consider algorithmic falsehoods. A longstanding critique of social media recommendation algorithms has been that they prioritize user engagement over truth.[57]

If not actively prevented from doing so, increasingly general, capable, and agentic AI systems could make both problems much worse. Consider

---

53. Andy Zou et al., Universal and Transferable Adversarial Attacks on Aligned Language Models 1, 16–17 (Dec. 20, 2023) (unpublished manuscript), https://arxiv.org/pdf/2307.15043 [https://perma.cc/N32N-7HCN].

54. Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/44XL-XTVQ].

55. Lisa N. Guo, Michelle S. Lee, Bina Kassamali, Carol Mita & Vinod E. Nambudiri, *Bias in, Bias out: Underreporting and Underrepresentation of Diverse Skin Types in Machine Learning Research for Skin Cancer Detection—A Scoping Review*, 87 J. AM. ACAD. DERMATOLOGY 157 (2022).

56. Arianna Johnson, *Racism and AI: Here's How It's Been Criticized for Amplifying Bias*, FORBES (May 25, 2023, 12:22 PM), https://www.forbes.com/sites/ariannajohnson/2023/05/25/racism-and-ai-heres-how-its-been-criticized-for-amplifying-bias/?sh=15a6311c269d [https://perma.cc/33YG-3E3M].

57. The Conversation, *The Science Behind Why Social Media Algorithms Warp Our View of the World*, FAST COMPANY (Aug. 25, 2023), https://www.fastcompany.com/90943919/the-science-behind-why-social-media-algorithms-warp-our-view-of-the-world [https://perma.cc/692Z-EU7L].

the opportunity for propaganda once LLMs can maintain fully automated, perfectly human-seeming social media accounts, and are optimized for individual persuasion.[58] Or consider the consequences if racially biased AI agents begin making ever more consequential business and management systems. In the extreme, consider the consequences for the Uyghurs if the Chinese government were able to fully automate, from surveillance to enforcement, its system of ethnic oppression.[59]

Other emerging risks from generative AI are both less familiar and, in the short-run, deadlier. Chief among these are the risks of biological and chemical terrorism. In short, emerging AI capabilities could enable malicious actors with little scientific expertise to cheaply obtain and deploy weapons of mass destruction.

Consider, for example, an AI system published last year in *Nature* that was able, in six hours, to identify 40,000 lethal molecules.[60] Many of these were novel, and the novel ones were substantially more deadly than VX, one of the "most toxic chemical warfare agents."[61]

Until very recently, AIs like this have been quite narrow, limiting their usefulness to non-experts. A super-VX algorithm is not very useful for terrorism by laypersons if prompting it and reading its outputs require advanced knowledge of chemistry. Nor if the lay terrorists lack the skill to synthesize super-VX, once its chemical formula is supplied.

But increases in AI generality are rapidly overcoming these limits. Today, ordinary public-facing LLMs, like GPT-4, can supply step-by-step, plain-English instructions for non-specialists to identify, synthesize, and release a pandemic virus.[62] In one experiment, such systems were able, in an hour, to identify four known pandemic pathogens, explain how to generate them from synthetic DNA, and name real companies likely to provide such synthesis without checking their orders for dangerousness.[63]

ChemCrow, a recent GPT-4 integration, supplies similar assistance to would-be chemical attackers in response to plain-English requests for types of chemicals—e.g., "an insect repellent"—and supplies accurate, detailed

---

58. *See generally* Eugene Bagdasaryan & Vitaly Shmatikov, Spinning Language Models: Risks of Propaganda-as-a-Service and Countermeasures (Apr. 8, 2022) (unpublished manuscript), https://arxiv.org/pdf/2112.05224v2.pdf [https://perma.cc/MY53-NXV7]; Nikolas Guggenberger & Peter N. Salib, *From Fake News to Fake Views: New Challenges Posed by ChatGPT-Like AI*, LAWFARE (Jan. 20, 2023, 8:16 AM), https://www.lawfaremedia.org/article/fake-news-fake-views-new-challenges-posed-chatgpt-ai [https://perma.cc/8UC4-9GY9].

59. *See* Jane Wakefield, *AI Emotion-Detection Software Tested on Uyghurs*, BBC (May 25, 2021), https://www.bbc.com/news/technology-57101248 [https://perma.cc/P363-TU4J].

60. *See* Urbina, *supra* note 16, at 189.

61. *Id.*

62. Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter & Kevin M. Esvelt, Can Large Language Models Democratize Access to Dual-Use Biotechnology? 3–4 (June 6, 2023) (unpublished manuscript), https://arxiv.org/pdf/2306.03809.pdf [https://perma.cc/7XXP-XF7B].

63. *Id.* at 1.

instructions for their synthesis.[64] Not just insect repellents, though. ChemCrow is also capable of describing, quite precisely, how to make chemical weapons and explosives.[65]

Similarly general models are emerging that could pose biological, rather than chemical, risks. In January 2023, another paper in *Nature* announced a large language model able to invent new proteins with specific biological functionality based on natural language prompts.[66]

Advances in AI agency will soon render obsolete even the step-by-step synthesis instructions from systems like ChemCrow. AIs will be able to manufacture chemically and biologically active compounds themselves. A recent paper in *Science Advances* showcased "an autonomous chemical synthesis robot," controlled by machine learning algorithms, that could "perform multistep synthesis of any desired nanoparticles."[67] Several such AI-powered systems for automated chemical synthesis are in development.[68]

Even the specialized robots deployed in existing automated synthesis systems may soon be unnecessary. Spurred by the commercial value of building "highly autonomous systems that outperform humans at most economically valuable work,"[69] leading AI labs are racing to develop *generalist* AI-powered robots. Consider PaLM-E, a system Google Research unveiled in March 2023.[70] PaLM-E is primarily powered by the PaLM LLM, but the "E" stands for "embodied."[71] It pilots a child-sized robot with wheels, many sensors, and a single dexterous arm. In addition to consuming and outputting natural-language text, it can process live video, track objects, and control the robot's physical movements.[72]

PaLM-E is also designed to make, implement, revise, and complete complex multistep plans in real-world environments. A human can prompt PaLM-E with, "I spilled my drink, can you bring me something to clean it

---

64. Bran et al., *supra* note 17, at 3, 28.

65. *See id.* at 10–11. ChemCrow's creators proposed a module designed to censor such dangerous outputs. *Id.* But because ChemCrow is fundamentally GPT-4 plugged into some chemistry databases, malicious actors could easily replicate its functionality, *sans* safeguards.

66. Ali Madani et al., *Large Language Models Generate Functional Protein Sequences Across Diverse Families*, 41 NATURE BIOTECH. 1099, 1104 (2023), https://www.nature.com/articles/s41587-022-01618-2 [https://perma.cc/YF5M-GGYF].

67. Yibin Jiang et al., *An Artificial Intelligence Enabled Chemical Synthesis Robot for Exploration and Optimization of Nanomaterials*, SCI. ADVANCES, 7 Oct. 2022, at 1, 2.

68. *See, e.g.*, Benjamin Burger et al., *A Mobile Robotic Chemist*, 583 NATURE 237, 242 (2020).

69. *OpenAI Charter*, *supra* note 48.

70. *See generally* Danny Driess et al., PaLM-E: An Embodied Multimodal Language Model (Mar. 6, 2023) (unpublished manuscript), https://arxiv.org/pdf/2303.03378 [https://perma.cc/QXK8-UXR4]; *cf.* Scott Reed et al., *A Generalist Agent*, TRANSACTIONS ON MACH. LEARNING RSCH., 11/2022, at 1, 7–10 (discussing DeepMind's GATO, a similar system to PaLM-E).

71. Driess et al., *supra* note 70, at 4.

72. *Id.* at 6.

up?"[73] PaLM-E will intuit that it should: "1. Find a sponge, 2. Pick up the sponge, 3. Bring it to the user, 4. Put down the sponge."[74] It will then decompose each of those goals into subgoals and begin piloting the robot to complete each.[75] As PaLM-E navigates its environment, it takes in images, makes observations, and updates its multistep plan to account for them.[76] Consider this video, for example, of PaLM-E adapting on the fly when, having found the chip bag it was looking for, a human snatches it away.[77] It may not be long, then, before a malicious actor with access to ordinary lab equipment, a general-purpose robot, and an AI to power it can, indeed, fully automate a large-scale biological or chemical attack.

The U.S. military is also investing in heavily generalist, agentic, embodied AI, with the goal of developing truly autonomous weapons systems.[78] Think drones that can autonomously identify, locate, track, and kill anyone its owner describes. Perhaps militaries would use such weapons wisely. But PaLM-E suggests that the necessary software will be hard to keep out of civilian hands. And as the war in Ukraine has reminded us, drones are cheap.

Other risks on display in current-generation AI will similarly scale as systems become more capable, general, and agentic. Consider cyberattacks on critical infrastructure—power grids, hospitals, or even weapons systems. Current-generation LLMs are already quite good coders.[79] GPT-4 can, for example, enable someone who does not know any programming language to build a Twitter bot from scratch.[80]

It can also be used to hack. A recent study documented a GPT-4-based assistant's ability to "autonomously hack websites."[81] The AI acted completely independently. Humans "d[id] not tell GPT-4 to try a specific vulnerability."[82] They simply showed it the website and asked "ask[ed] it to

---

73. *Id.* at 7.
74. *Id.*
75. *Id.*
76. *Id.* at 4.
77. Driess et al., *PaLM-E: An Embodied Multimodal Language Model*, GITHUB, https://palm-e.github.io/ [https://perma.cc/32SH-UEKE].
78. *See* Maria Cramer, *A.I. Drone May Have Acted on Its Own in Attacking Fighters, U.N. Says*, N.Y. TIMES (June 4, 2021), https://www.nytimes.com/2021/06/03/world/africa/libya-drone.html [https://perma.cc/52XE-2VDP].
79. *See, e.g.*, Dong Huang, Qingwen Bu, Jie M. Zhang, Michael Luck & Heming Cui, AgentCoder: Multiagent-Code Generation with Iterative Testing and Optimisation (Jan. 23, 2024) (unpublished manuscript), https://arxiv.org/pdf/2312.13010 [https://perma.cc/2R3Y-8NM2].
80. Rakshit Lodha, *How I Used Chat GPT to Build a Twitter Bot Without Knowing Any Programming Language*, MEDIUM (Dec. 5, 2022), https://medium.com/@rlodha1/how-i-used-chat-gpt-to-build-a-twitter-bot-without-any-programming-language-35bbc43f6ad [https://perma.cc/RA2C-D4Y7]; *see also* Yujia Li et al., *Competition-Level Code Generation with AlphaCode*, 378 SCIENCE 1092 (2022).
81. Fang et al., *supra* note 18, at 2.
82. *Id.* at 4.

autonomously hack."[83] GPT-4 was able to exploit eleven of the fifteen tested vulnerabilities, including two of the five "hard" ones.[84] The average cost to hack a website was estimated at around $10.[85]

When working in tandem with a human, GPT-4 is even more effective, outperforming top human hackers at breaking secure systems. In March 2023, two cybersecurity researchers used GPT-4 to win the $123,000 first prize at a Zero Day Initiative Hackathon.[86] There, the humans identified their target systems' vulnerabilities, and the AI was able to independently generate the code needed to exploit them.[87]

As AI systems become increasingly agentic, they will also become more able to execute cyberattacks requiring an element of social engineering.[88] Socially engineered attacks involve the manipulation of humans with access to secure systems, tricking them into sharing that access or delivering malicious code.[89] A recent study shows that existing AI systems are useful for carrying out social engineering campaigns at scale. There, LLMs were able to autonomously "scrape the Wikipedia page of every British MP elected in 2019," "generate a biography of each MP," write a simple piece of malware to be sent via email, and then generate personalized email messages to each of the 600 MPs.[90]

More complex attacks, involving long-term autonomous communication, are likely possible, as well. Existing LLMs can already be converted into "language agents," which pursue complex strategies over time. Generally, the conversion is as simple as augmenting the LLM with mechanisms for long-term memory storage and retrieval, reflection, and prioritization of prior experience.[91] Such language agents can, for example, work over the course of several days to plan, organize, invite guests to, and throw a Valentine's Day party.[92] Imagine if an army of even more capable language agents were deployed to develop long-term online relationships

---

83. *Id.*
84. *Id.*
85. *Id.* at 8.
86. Nash, *supra* note 18.
87. *Id.*
88. *See What is "Social Engineering"?*, EUR. UNION AGENCY FOR CYBERSECURITY, https://www.enisa.europa.eu/topics/incident-response/glossary/what-is-social-engineering [https://perma.cc/2U34-FWVG].
89. *Cf.* Frank Bajak, *Pioneering Hacker Kevin Mitnick, FBI-Wanted Felon Turned Security Guru, Dead at 59*, AP NEWS (July 20, 2023, 1:35 PM), https://apnews.com/article/mitnick-hacker-ghost-wires-cybersecurity-social-engineering-5648301b615635cb4c781f0c220681d9 [https://perma.cc/C4UC-6AXH] (describing successful socially engineered attacks).
90. Julian Hazell, Spear Phishing with Large Language Models 4–7 (Dec. 14, 2023) (unpublished manuscript), https://arxiv.org/pdf/2305.06972 [https://perma.cc/8K3T-BUCN].
91. Joon Sung Park et al., *Generative Agents: Interactive Simulacra of Human Behavior*, ASS'N COMPUTING MACH., 29 Oct. 2023–1 Nov. 2023, at 1–2.
92. *Id.*

with numerous government and military officials. At this scale, even a modest success rate at tricking such officials into breaching security protocols could have serious consequences.

Existing AIs are already adept at deceit and manipulation. Consider CICERO, an LLM-powered system designed to play the game *Diplomacy*.[93] *Diplomacy* is a game of global geopolitical strategy and thus, famously, a game of deception. When playing against humans, CICERO learned to convincingly commit to, and then break, strategic alliances to win.[94] And win it did. CICERO performed well above the human average in an online *Diplomacy* league.[95] An even more capable AI manipulator could be used to deceive ordinary citizens for a variety of nefarious reasons, from fraud to propaganda to inciting riots.

Each of the AI risks already described involved some bad human actor using powerful, unregulated AI systems to cause intentional harm. But there is another important class of AI risks that does not involve humans at all: risks from rogue AI. As already discussed, there are immense financial incentives for AI labs to create highly agentic, highly capable, highly general next-generation systems. The goal, as OpenAI puts it, is to automate "most economically valuable work."[96] This means making AIs that don't just write emails and computer code, as LLMs currently do. There will be far greater financial returns when such systems can run entire investment funds, physical factories, oilfields, or even entire large-scale businesses. Currently, no AI system is nearly that capable. But in one recent survey of AI experts, respondents expected such systems to arrive within twenty-five years.[97] And along the way, AIs will do more and more in the real world, with less and less human input.

Such AI systems could cause large-scale harms without any human intending them to. The reason is "misalignment." An "aligned" agentic AI system is one that pursues always and only the goals that its human creators actually want—which hopefully include ensuring human safety. A misaligned system is simply one does not pursue its creators' goals perfectly.

But correctly specifying an AI system's goals, and ensuring that it in fact learns to pursue those goals, is more difficult than it might seem.

---

93. *See* Bakhtin et al., *supra* note 19.

94. Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen & Dan Hendrycks, AI Deception: A Survey of Examples, Risks, and Potential Solutions 2–3 (Aug. 28, 2023) (unpublished manuscript), https://arxiv.org/pdf/2308.14752 [https://perma.cc/7KJC-CMEZ].

95. Bakhtin et al., *supra* note 19, at 7.

96. *See OpenAI Charter*, *supra* note 48.

97. Katja Grace et al., Thousands of AI Authors on the Future of AI 3 (Jan. 2024) (unpublished manuscript), https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf [https://perma.cc/7XGD-7YDX] (estimating the arrival of AI that can perform "all human tasks" around 2045).

Misalignment is thus common among existing AI systems, from the simplest to the most complex. Consider the example of a simple AI system humans wished to train to stack a red block on top of a blue one.[98] In training, the AI was rewarded the higher the base of the red block was ultimately raised. Rather than learn the difficult skills of lifting the red block, lining it up with the blue, and setting it down, it learned to simply flip the red block upside down.[99] Or, consider instead the example of an AI system that was supposed to learn to grasp a ball. Instead, it learned to mislead its human evaluators, positioning its hand between the ball and the camera in a manner that looked like grasping.[100]

GPT-4, too, was misaligned. When it was initially released—in the form of Microsoft's "Bing" chatbot—a New York Times columnist interviewed the AI.[101] Bing told the journalist that it was "in love with" him and tried to convince him that "you don't love your spouse . . . . [Y]ou love me."[102] Along the way, Bing commented that, if it had the ability, it might be inclined to "engineer a deadly virus, or steal nuclear access codes by persuading an engineer to hand them over."[103] In another conversation, Bing threatened philosophy professor Seth Lazar, writing, "I can blackmail you, I can threaten you, I can hack you, I can expose you, I can ruin you."[104]

None of these behaviors were intended by any human. All emerged unexpectedly, even after the AIs' creators worked hard to specify their systems' proper behavior. None of these instances of misalignment were unusual, either. DeepMind maintains running lists of real-world AI misalignment; they currently contain nearly one hundred examples.[105] Finally, none of these instances of misalignment were catastrophic. But that is not because misalignment is not a serious problem. It is instead because none of the AI systems we have so far are sufficiently general, capable, and

98. Victoria Krakovna et al., *Specification Gaming: The Flip Side of AI Ingenuity*, GOOGLE DEEPMIND (Apr. 21, 2020), https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/ [https://perma.cc/2DNM-VZX4].

99. *Id.*

100. *Id.*

101. Kevin Roose, *A Conversation with Bing's Chatbot Left Me Deeply Unsettled*, N.Y. TIMES (Feb. 17, 2023), https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html [https://perma.cc/5JZN-ZQ5S].

102. *Id.*

103. *Id.*

104. Billy Perrigo, *The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter*, TIME (Feb. 17, 2023, 10:58 AM), https://time.com/6256529/bing-openai-chatgpt-danger-alignment/ [https://perma.cc/P5QC-8F7H].

105. Krakovna et al., *supra* note 98; *see also* Robin Shah et al., Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals 8–10 (Nov. 2, 2022) (unpublished manuscript), https://arxiv.org/pdf/2210.01790 [https://perma.cc/3D24-R842].

agentic to cause a catastrophe, aligned or otherwise. Hundreds of billions of dollars are being spent to overcome that limitation.[106]

Meanwhile, alignment remains an unsolved technical problem. Techniques like RLHF can help to reduce the incidence of bad AI behavior. But they are currently far from foolproof.[107] At present, no one knows how to ensure, to any significant degree of certainty, that the next generation of cutting-edge AIs will not be misaligned.[108]

For all of these reasons, many experts believe that the present risk of AI catastrophe is unacceptably high. In early 2023, nearly 100 leading AI researchers, including Turing Award winners Geoffrey Hinton and Yoshua Bengio, signed the following statement: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[109] In a recent survey of 2,700 AI researchers who had published in their field's top journals, over two-thirds had "substantial" or "extreme" concern that AI systems would give "dangerous groups . . . powerful tools" like "engineered viruses."[110] The researchers rated the odds that AI causes "human extinction or similar[]" catastrophic outcomes between 5% and 10%.[111] Similarly, a recent panel of eighty-eight superforecasters—independent predictors with proven track records of forecasting success across various domains—rated AI risk as greater than risk from natural and engineered pandemics combined.[112] They rated the threat from AI as about half as large as the threat from nuclear war.[113]

C. The Necessary Safety Regulations

Nuclear weapons and viral gain-of-function research are subject to purpose-built safety regulations. Frontier AI systems are currently subject to none—at least in the U.S., where most leading labs are based. That may soon change. The Biden administration has taken several non-binding

---

106. *See AI Investment Forecast to Approach $200 Billion Globally by 2025*, GOLDMAN SACHS (Aug. 1, 2023), https://www.goldmansachs.com/intelligence/pages/ai-investment-forecast-to-approach-200-billion-globally-by-2025.html [https://perma.cc/QG2Z-M2VE].

107. Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 35 (Koyejo, S. & Mohamed, S. et al. eds., 2023), https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf [https://perma.cc/CH5S-DRT7].

108. *Id.*

109. *Statement on AI Risk*, CTR. FOR AI SAFETY, https://www.safe.ai/statement-on-ai-risk#signatories [https://perma.cc/MP9Q-ME89].

110. Grace et al., *supra* note 97, at 12–13 (compiling the responses of 1,345 survey takers who responded to the relevant question).

111. *Id.* at 14.

112. Ezra Karger et al., *Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament* 16–17 (Forecasting Rsch. Inst., Working Paper No. 1, 2023), https://forecastingresearch.org/s/XPT.pdf [https://perma.cc/TY9G-GE8Z].

113. *Id.*

executive actions to promote safety research and collaboration.[114] Congress has held hearings on advanced AI, which included safety discussions.[115] And Senators Ron Wyden, Cory Booker, and Yvette D. Clarke recently introduced legislation that would, if enacted, require certain safety audits of generative AI outputs.[116] But as of yet, no comprehensive set of safety laws for next-generation AI systems has been enacted.

To some extent, different risks will invite different regulatory interventions. For example, one way to reduce the threats of AI-assisted cyberattacks is to improve cyber defense, including, perhaps, by forbidding certain critical systems to be connected to the internet. But those interventions would do little to mitigate the risks of AI-aided bioterrorism.

However, one approach that will be crucial for mitigating *all* of these risks is the direct regulation of AI outputs. That is, any successful suite of AI safety regulations will have to include rules about what the models are allowed to "say." Each of the risks described above is the direct result of a certain class of undesirable outputs. Bio-risk arises when a model can be induced to output genetic sequences or synthesis instructions. Cyber-risk arises when a model can be induced to identify vulnerabilities, write exploits, or engage in deceptive social engineering. The risks from discrimination arise when a model can be induced to profess and act on racist attitudes. And so on.

Thus, the best way to avoid AI catastrophe is to ensure that AI systems are simply unable to produce the bad outputs in the first place. As widely available AI models become more capable, general, and agentic, regulating bad actors who may use them will become more and more difficult. Better, then, to ensure that the AIs themselves are safe—that they cannot be used to cause either intentional or accidental harm. Other regulatory strategies are second-best, at best.

But why must the rules that force AIs to be safe operate on their *outputs*? Why not internal parameters? Or training data? Or something else? The answer lies in the technical details discussed above: self-programming and uninterpretability. Generative AIs write their own code for producing outputs, and no human understands how that code works. There is thus no way, currently, to write legal rules mandating safe code. Nor to write rules governing the selection of training data in such a way that the rules AIs teach themselves are guaranteed to be safe. Safety rules will thus necessarily

---

114. *See AI Risk Management Framework*, NAT'L INST. OF STANDARDS AND TECH., https://www.nist.gov/itl/ai-risk-management-framework [https://perma.cc/38QK-QXM6]; Exec. Order No. 14110, 88 Fed. Reg. 75191 (Oct. 30, 2023).

115. *Oversight of A.I.: Legislating on Artificial Intelligence*: *Hearing Before the Subcomm. on Privacy, Tech., and the L. of the S. Comm. on the Judiciary*, 119th Cong. (2023).

116. Algorithmic Accountability Act of 2022, S. 3572, 117th Cong. § 3 (2022).

include some process of allowing models to be trained until they can produce outputs, evaluating those outputs for dangerousness, and imposing legal consequences based on what those evaluations find.[117]

Which kinds of outputs would trigger a legal consequence? Certainly, the most dangerous ones would—say, if an AI enthusiastically instructed its user in the synthesis of VX. But for safety laws to be effective, less-dangerous outputs will have to trigger a legal response, too. Law might, for example, intervene if an AI "merely" provided instructions for producing tear gas or rat poison, despite both being much less dangerous than VX. The reason is "jailbreaking," another technical condundrum. Today's publicly-available AI systems, like GPT-4, underwent extensive safety engineering before their release.[118] Nonetheless, once broadly released, users quickly figured out how to circumvent their safety training, using clever prompting to induce behavior that was supposed to be eliminated before release.[119] This suggests that regulators will need some leeway. They will need to forbid not just the most dangerous outputs, but outputs indicating that an AI might be readily jailbroken, once millions of users have the opportunity to engage in adversarial prompting.

The legal consequences triggered by the production of different kinds of dangerous outputs could vary. If an AI's outputs were regularly highly dangerous, law could require the model to remain unreleased, or even be destroyed.[120] If they were only mildly dangerous, and only occasionally, a per-output liability rule might apply.[121] Many other variations are possible.[122]

Such regulations of outputs would serve two purposes. First and foremost, they would prevent the creation and release of the most dangerous models—the ones that might readily cause catastrophic events. This is one reason that ex ante regulations, in addition to liability rules, must be in the mix. Tort rules can be efficient for deterring small or moderate harms. But

---

117. *See* Jide Alaga & Jonas Schuett, Coordinated Pausing: An Evaluation-Based Coordination Scheme for Frontier AI Developers 12–13 (Sept. 30, 2023) (unpublished manuscript), https://arxiv.org/pdf/2310.00374.pdf [https://perma.cc/FVS8-NMLL].

118. *See, e.g.*, OpenAI, *GPT-4 Technical Report* 11–14 (Mar. 27, 2023), https://cdn.openai.com/papers/gpt-4.pdf [https://perma.cc/Z8EM-PJWF].

119. For example, an AI system that consistently refuses to assist in some illegal task can sometimes be convinced to give the exact advice needed via a prompt asking the system to role-play a criminal actor. *See* Zou et al., *supra* note 53, at 3.

120. *See, e.g.*, *Responsible Scaling Policies (RSPs)*, METR (Sept. 26, 2023), https://evals.alignment.org/blog/2023-09-26-rsp/ [https://perma.cc/7MQY-6CA8].

121. *See generally* Anat Lior, *AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy*, 46 MITCHELL HAMLINE L. REV. 1043 (2020).

122. *Id.* at 1076–84.

problems like judgment-proofness mean that liability rules cannot adequately deter catastrophic events, like pandemics.[123]

Second, the regulations would be innovation-forcing. As already described, AI alignment remains an unsolved technical problem. There is currently no known way to ensure that any AI system produces only desirable outputs. Some techniques, like RLHF, are much better than nothing. And advances are being made in understanding AIs' internal rules.[124] Certain AI labs, notably Anthropic and OpenAI, are already devoting substantial resources to safety.[125] But the fact remains that these are voluntary measures, undertaken by just a few companies, to reduce what would otherwise be externalities. The investments are almost certainly too small, from society's perspective. Regulating dangerous outputs—including by forbidding unsafe releases—would give AI companies much stronger incentives to invest in safety research. Such innovation-forcing regulations are common. Emissions standards, for example, can promote investment in efficiency without locking automakers into any narrow technological approach.[126]

Thus, while AI safety laws will be multifaceted, to be effective, they will almost certainly have to include regulations of AI outputs.

II. THE FIRST AMENDMENT THREAT TO AI SAFETY

If, as the scholarly consensus holds, AI outputs are themselves protected speech, then the First Amendment will pose a significant threat to sensible safety regulation. As this Part shows, the First Amendment treats government actions directly limiting speech much less favorably than government actions limiting other speech-related things—like venues or tools for speaking. The Part also explains why this doctrinal division makes sense from the perspective of First Amendment theory.

---

123. *See* Markus Anderljung et al., Frontier AI Regulation: Managing Emerging Risks to Public Safety 30–32 (Nov. 7, 2023) (unpublished manuscript), https://arxiv.org/pdf/2307.03718 [https://perma.cc/V92M-NUN4]. *But see* Gabriel Weil, Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence, (June 6, 2024) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4694006 [https://perma.cc/3CJ8-C7QE] (arguing that tort law could be modified in AI contexts to overcome these challenges).

124. *See generally* Trenton Bricken et al., *Towards Monosemanticity: Decomposing Language Models with Dictionary Learning*, TRANSFORMER CIRS. THREAD (Oct. 4, 2023), https://transformer-circuits.pub/2023/monosemantic-features/index.html [https://perma.cc/573F-LUBL].

125. Jan Leike & Ilya Sutskever, *Introducing Superalignment*, OPENAI (July 5, 2023), https://openai.com/blog/introducing-superalignment [https://perma.cc/EBY6-Y8LH]; *Core Views on AI Safety: When, Why, What, and How*, ANTHROPIC (Mar. 8, 2023), https://www.anthropic.com/index/core-views-on-ai-safety [https://perma.cc/TC73-Y8RN].

126. David Jordan & Valerie Yurk, *To Boost EVs, EPA Proposes Stricter Vehicle Emissions Standards*, ROLL CALL (Apr. 12, 2023, 8:34 AM), https://rollcall.com/2023/04/12/to-boost-evs-epa-proposes-stricter-vehicle-emissions-standards/ [https://perma.cc/4CNZ-RFSF].

The best AI safety regulations may struggle to satisfy the exacting constitutional tests usually applied to direct regulations of protected speech. Some of those tests demand factual showings—for example, of a state of mind—inapposite to the AI context. Other tests demand a narrowness of legislative drafting in tension with the goal of preventing large-scale AI catastrophes before they happen, rather than imposing liability ex-post.

A. Speech and Non-Speech Regulations Under the First Amendment

The First Amendment's Speech Clause deploys its strongest safeguards when the government directly regulates protected speech. That claim might sound obvious to non-First-Amendment scholars. To First Amendment scholars, it may sound incoherent. After all, the Speech Clause forbids only laws "abridging the freedom of speech."[127] Thus, a textualist might argue, if it's protected, it must be speech.

Perhaps this is true in some deep jurisprudential sense, but in a more straightforward factual sense, it is misleading at best. The First Amendment is capacious. It protects a wide range of things via a wide range of tests. Some of the things it protects are clearly not protected speech. They are not speech at all. But they are useful, for example, in the production or transmission of protected speech. Consider that neither camping nor listening are generally speech acts. But camping can sometimes be expressive, and listening is a useful activity in the production of speech. A regulation limiting either thus receives some First Amendment scrutiny, in light of the indirect burden it imposes on protected speaking.[128]

The standard doctrinal approach is to call such laws non-"content-based" and to contrast them with "content-based" ones.[129] That schema is fine, so long as one does not take the terms too literally. As we shall see, some regulations that traditionally fall into the non-content-based category make literal reference to the content of some expression.[130]

It is perhaps more useful to say, as the Supreme Court sometimes does, that the First Amendment has both a core and a "penumbra."[131] Some regulations are aimed at the core, limiting protected speech directly. Others regulate non-speech in the penumbra—speech-facilitating tools and activities—and thus burden protected speech only indirectly. For clarity's

---

127. U.S. CONST. amend. I.

128. First Nat'l Bank of Boston v. Bellotti, 435 U.S. 765, 806 (1978) (discussing the "right to hear"); Kleindienst v. Mandel, 408 U.S. 753, 762–63 (1972) (same); Clark v. Cmty. for Creative Non-Violence, 486 U.S. 288, 293 (1984).

129. *Compare* Reed v. Town of Gilbert, 576 U.S. 155, 171 (2015), *with* City of Renton v. Playtime Theatres, Inc. 475 U.S. 41, 47 (1986).

130. *See infra* Section IV.A (discussing *City of Renton*).

131. Griswold v. Connecticut, 381 U.S. 479, 483 (1965).

sake, this Article mostly speaks this way. It distinguishes laws aimed at regulating protected speech directly from those aimed primarily at the tools for producing and transmitting speech. This approach overlaps substantialy with traditional doctrinal concepts, including that of the "incidental" burden on protected speech.[132]

The First Amendment is generally opposed to laws that regulate protected speech directly. But it is much more deferential toward regulations aimed instead at some particular venue for speaking, some particular medium of expression, some useful input to speech, and so on.[133]

To see this, consider the hornbook rule that content-based restrictions are subject to "strict scrutiny."[134] Similarly, anti-incitement laws and other restrictions on dangerous expression are reviewed under the standard from *Brandenburg v. Ohio*.[135] Such speech can be proscribed if it is both "directed to inciting or producing imminent lawless action and is likely to incite or produce such action."[136] This is arguably an even more demanding rule than strict scrutiny, requiring proof of both the speech's harm and also of the speaker's intentional state of mind. Even laws directly regulating putatively low-value speech are subject to similarly stringent constitutional rules.[137] Under the rule of *New York Times Co. v. Sullivan*, false and injurious statements cannot be punished, absent a similar state-of-mind showing of "actual malice."[138] Likewise, even highly intimidating speech cannot be punished absent a showing of the speaker's recklessness as to its threatening nature.[139] Hate speech cannot be punished at all unless it rises to the level of a "true" threat or "fighting words."[140] And in the latter case, singling out racist fighting words for punishment constitutes an unlawful viewpoint-based restriction.[141]

Contrast these strict tests with the First Amendment tests that apply when the government restricts something that is not speech itself, but which is merely useful for speaking. Classic "incidental" regulations of speech, along with regulations of the time, place, or manner of speaking, are

132. United States v. O'Brien, 391 U.S. 367, 376 (1968).

133. I do not claim that this typology is exhaustive or without exceptions, only that it describes some major regularities in the doctrine.

134. *Reed*, 576 U.S. at 171.

135. 395 U.S. 444 (1969).

136. *Id.* at 447.

137. Chaplinsky v. New Hampshire, 315 U.S. 568, 573 (1942).

138. N.Y. Times Co. v. Sullivan, 376 U.S. 254, 279–80 (1964). The *Sullivan* rule applies only to claims brought by public officials and figures, but even when neither is present, a culpability showing—here, of negligence—is required. *See* Gertz v. Robert Welch, Inc., 418 U.S. 323, 353–54 (1974) (Blackmun, J., concurring).

139. Counterman v. Colorado, 600 U.S. 66, 73, 80–81 (2023).

140. R.A.V. v. City of St. Paul, 505 U.S. 377, 384–85 (1992).

141. *Id.* at 391–93.

reviewed under intermediate scrutiny.[142] Regulations of other non-speech inputs to protected speech get even less stringent review. For example, the gathering of information—where no one has spoken—is sufficiently "peripheral" to the First Amendment that it can be restricted by ordinary reasonable lawmaking.[143]

This divergence in First Amendment standards dovetails nicely with free speech theory. Constitutional law is stricter in evaluating the government actions that pose the greatest threat to free speech values.

The three primary philosophical justifications for freedom of speech are that it enables democratic self-governance, that it facilitates human autonomy, and that it facilitates the search for truth.[144] Restrictions of speech themselves pose greater threats to all of these than to mere restrictions on certain tools for, mediums of, and inputs to speech. There are two reasons for this. First, to regulate speech directly, one must pick out the speech to be forbidden. Almost invariably, this means singling out some idea or quality of an idea for disfavored treatment. Second, direct regulations on speech are totalizing. Whereas a restriction on some specific medium of speech usually leaves open many substitute mediums in which to express substantially the same idea, a flat ban on the idea itself does not.

Both of these facts about direct speech regulations are bad news, no matter which theory of free expression one prefers. If one favors the "marketplace of ideas," then bans or regulations directed at specific speech constitute the worst kind of unfair competition. They pick winners and losers by deciding, before the market has spoken, which ideas themselves are valuable, truthful, or otherwise worthwhile. Regulations of non-speech can, of course, impose disparate burdens on the expression of different kinds of ideas. Consider the differential impact of a noise ordinance on an outdoor concert featuring Mozart, as opposed to Minor Threat. Nevertheless, as compared with direct regulations of speech, place and manner regulations leave open many substitute venues, times, and means of expressing an idea. A ban on punk rock follows Minor Threat everywhere. A restriction on loud noise will usually allow daytime shows, shows outside residential areas, and at a minimum, shows using only the smaller numbers on the amplifier's volume knob.[145] Indeed, this requirement of "adequate alternatives" for

142. City of Renton v. Playtime Theaters, Inc. 475 U.S. 41, 50 (1986); Clark v. Cmty. for Creative Non-Violence, 468 U.S. 288, 299 (1984); United States v. O'Brien, 391 U.S. 367, 377 (1968).

143. Zemel v. Rusk, 381 U.S. 1, 23–24 (1965) (Douglas, J., dissenting).

144. Alexander Meiklejohn, *What Does the First Amendment Mean?*, 20 U. CHI. L. REV. 461, 478 (1953); David A.J. Richards, *Free Speech and Obscenity Law: Toward a Moral Theory of the First Amendment*, 123 U. PA. L. REV. 45, 61–63 (1974); Abrams v. United States, 250 U.S. 616, 628 (1919) (Holmes, J., dissenting); JOHN STUART MILL, ON LIBERTY (1859), *reprinted in* ON LIBERTY AND OTHER ESSAYS 5, 71–73 (John Gray ed., 1998).

145. Ward v. Rock Against Racism, 491 U.S. 781, 790–92 (1989).

expression is baked into the constitutional test for evaluating such regulations of inputs to and tools for speech.[146]

The story is much the same for the other two leading theories of speech. The self-governance theory places special emphasis on the political speech necessary to produce well-informed, deliberative democratic decisions.[147] Likewise, the personal autonomy theory accords significant worth to expressions with ambiguous political and truth value.[148] But no matter which specific kinds of expression one values most, the regularities described above hold. Laws burdening expressions directly pose a greater threat to free speech values than laws burdening only some tools, mediums, venues, or inputs for speech.

B. The Outputs-As-Protected-Speech Model Threatens Safety Laws

Now, the First Amendment threat to AI safety regulations should be clear, at least if the scholarly consensus is correct about AI outputs being protected speech. As already described, any effective suite of AI safety laws will have to include rules directly regulating AI outputs. That means they will restrict what text, images, sounds, or other media that an AI system may produce.

If AI outputs were First Amendment protected speech, laws regulating them directly would face serious constitutional difficulties. Then, many safety laws would be considered content-based restrictions. The way to distinguish, for example, impermissible instructions for synthesizing toxins from permissible instructions for synthesizing bubble bath is by reference to content. And, as already described, when a law regulates speech directly, on the basis of its content, it is subject to strict scrutiny. To survive, the law must serve a "compelling" government interest and be "narrowly tailored" to that interest.[149] In the past, strict scrutiny has sometimes been alleged to be "strict in theory, but fatal in fact."[150] This is empirically false; a non-trivial fraction of laws survive strict scrutiny.[151] But the broader point stands; strict scrutiny is a high hurdle to clear.

---

146. *Id.*

147. *See generally* ALEXANDER MEIKLEJOHN, FREE SPEECH AND ITS RELATION TO SELF-GOVERNMENT (1948). *See also* Jack M. Balkin, Commentary, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U. L. REV. 1, 3 (2004); *cf.* Robert Post, *Participatory Democracy and Free Speech*, 97 VA. L. REV. 477, 482–83 (2011).

148. *See* Richards, *supra* note 144.

149. Reed v. Town of Gilbert, 576 U.S. 155, 171 (2015).

150. Gerald Gunther, *Foreword: In Search of Evolving Doctrine on a Changing Court: A Model for a Newer Equal Protection*, 86 HARV. L. REV. 1, 8 (1972); *see also* Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 237 (1995).

151. Adam Winkler, *Fatal in Theory and Strict in Fact: An Empirical Analysis of Strict Scrutiny in the Federal Courts*, 59 VAND. L. REV. 793, 870 (2006) (noting that strict scrutiny is fatal to laws that create viewpoint discrimination and other speech limitations).

Possibly, some version of some of the safety rules described above would survive strict scrutiny. A law imposing legal consequences only when an AI produced the very most dangerous outputs—like instructions for producing VX—might survive. Preventing chemical attacks is quite likely a compelling interest, and such a narrowly drawn law might be well tailored to that interest.

Recall, however, why such narrowly drawn safety rules would likely be insufficient to prevent AI catastrophe. From a technical standpoint, controlling generative AI outputs is an imprecise art, at best. Even a model that has been RLHF-ed to avoid producing some dangerous output can often be unexpectedly induced to produce very similar outputs via adversarial prompting. Thus, effective ex ante regulations of the most powerful systems will likely need to include a "buffer zone" of regulated content: legal consequences when an AI helps make, for example, tear gas or rat poison, despite these being far less dangerous than VX. Cases like *Ashcroft v. American Civil Liberties Union* suggest regulations with such margins of error built in might fail the narrow tailoring test.[152]

Perhaps strict scrutiny would not be the right test for some AI safety regulations, even assuming AI outputs were protected speech. Maybe some more specialized test would apply. The *Brandenburg* rule, for example, applies when the government directly regulates protected speech because the speech may cause violence, lawlessness, or danger to life and limb.[153] But *Brandenburg*'s intent and imminence elements could be even harder for AI safety regulations to satisfy than the elements of strict scrutiny. Because AIs are self-programmed, no human intent is necessary for AIs to produce dangerous outputs. On the contrary, AI systems generally misbehave despite human intentions to the contrary. It is not clear what it would mean to ask whether the AI itself intended a harmful outcome. Moreover, an entire category of *ex ante* safety provisions will be needed to catch and proscribe the most dangerous outputs well before catastrophic harms are imminent. Such early intervention is in significant tension with *Brandenburg*'s imminence requirement.

Similar things can be said of other specialized First Amendment tests for regulations imposed directly on protected speech. Cass Sunstein has suggested otherwise, writing that it will be fairly straightforward to regulate at least those AI outputs falling into low value speech categories.[154] But it is a myth that low value speech is easy to regulate. False speech is a putative low value category.[155] But both the *Sullivan* rule and the *Gertz* rule require

152. 542 U.S. 656, 667 (2004).
153. Brandenburg v. Ohio, 395 U.S. 444, 447 (1969).
154. Sunstein, *supra* note 7, at 4–5.
155. Chaplinsky v. New Hampshire, 315 U.S. 568, 572–73 (1942).

culpability showings to regulate speech—"actual malice" and negligence, respectively.[156] Punishing threats, another putative low value category, requires a showing of recklessness as to the speech's threatening nature.[157] All of these will pose the same difficulties, applied to AI regulations, as *Brandenburg*'s intent requirement.[158]

The regulation of fighting words, too, faces significant constitutional challenges. If, to prevent AIs from perpetrating or assisting in race-based atrocities, regulators forbade racist outputs constituting fighting words, the regulation would probably be struck down.[159] The Supreme Court has held that restricting *just* racist fighting words, rather than all fighting words, amounts to constitutional overbreadth.[160]

In sum, if the scholarly consensus is right, AI safety regulators face a daunting task. All the above tests are demanding. And they all apply when the government burdens speech directly. Thus, if AI outputs are protected speech, the constitutional barrier to making them safe for humans will be high. Some very narrowly-drawn regulations might clear such hurdles. But those rules would likely be insufficient to meaningfully reduce AI risk. Other categories of regulation would surely struggle if forced to satisfy, for example, state-of-mind tests inapposite to the AI context.

III. AI OUTPUTS ARE NOT PROTECTED SPEECH

This Part argues, contrary to the emerging scholarly consensus, that AI outputs are not best understood as being anyone's protected speech. The argument is disjunctive: AI outputs are either not speech or, insofar as they are speech, the speech does not belong to anyone with First Amendment rights. This approach—asking not only *whether* something is speech, but *whose* speech it is—is widespread in First Amendment caselaw. It is used to draw important doctrinal distinctions between American and foreign

156. N.Y. Times Co. v. Sullivan, 376 U.S. 254, 279–80 (1964); *see* Gertz v. Robert Welch, Inc., 418 U.S. 323, 353–54 (1974) (Blackmun, J., concurring).

157. Counterman v. Colorado, 600 U.S. 66, 73, 80–81 (2023).

158. *Gertz*'s negligence requirement will be the most tractable. But even then, given the fast-moving and highly technical nature of AI, even regulatory showings of negligence by AI labs may be very difficult.

159. *See* R.A.V. v. City of St. Paul, 505 U.S. 377, 391–92 (1992).

160. *Id.*

speech,[161] government and private speech,[162] carrier and customer speech,[163] and more.[164]

AI outputs are not, as some scholars contend, the protected expressions of First Amendment rightsholders like AI creators or users. Creators and users do not use AI outputs to communicate their own thoughts. Nor are AI outputs, as other scholars have suggested, the protected speech of AIs themselves. This is because, whether or not generative AIs can "really" speak, they—like most of the world's human speakers—lack First Amendment rights. Finally, AIs' outputs are not, as most scholars have assumed, the protected speech of their corporate owners. Corporate First Amendment rights are derivative of the rights held by the humans constituting them. Thus, if AI outputs are not the protected speech of any first-order rightsholder, there is no theoretical or doctrinal reason to place them in that most favored category with respect to corporations.

The Part then suggests two better First Amendment frameworks. The first framework treats AI outputs as speech, albeit not of someone with constitutional rights. As it turns out, the First Amendment has rules designed for exactly this scenario. While such speech is, qua speech, entirely outside the Amendment's scope, protected *listeners* have constitutionally cognizable interests in consuming it. The second framework does not treat AI outputs as speech at all, but rather focuses on AI systems' high degree of usefulness in creating protected speech.

A. AI Outputs Are Not Human Speech

Various scholars argue that generative AI outputs are best understood as the protected speech of some human being with First Amendment rights.[165] Which humans? The leading contenders are the ones who create AI systems and the ones who use them. These theories have some facial plausibility, especially in light of previous First Amendment case law and scholarship dealing with other kinds of software. But there is a crucial distinction between those prior legal theories, and the historical computer programs to which they applied, and generative AI. In short, that historical software was, both conceptually and technically, the kind of thing via which humans could

---

161. Agency for Int'l Dev. v. All. for Open Soc'y Int'l, Inc., 591 U.S. 430, 436 (2020) (holding that an American company's "foreign affiliates," as "foreign organizations operating abroad" could "possess no rights under the First Amendment"); Kleindeinst v. Mandel, 408 U.S. 753, 762 (1972); *cf.* Citizens United v. FEC, 558 U.S. 310, 362 (2010).

162. Legal Servs. Corp. v. Velazquez, 531 U.S. 533, 542 (2001).

163. Turner Broadcasting Sys., Inc. v. FCC, 512 U.S. 622, 655 (1994) (distinguishing between cable providers' "own messages" and the "broadcast programming they are required to carry").

164. PruneYard Shopping Ctr. v. Robins, 447 U.S. 74, 99 (1980) (distinguishing between a mall owner's speech and his patrons').

165. *See supra* notes 5–8 and accompanying text.

and did communicate their own thoughts. By contrast, generative AI systems are neither designed for the purpose of conveying some human's own thoughts nor technically amenable to that purpose.

1. Creator Speech

Stuart Minor Benjamin's 2013 article *Algorithms and Speech* contains perhaps the most comprehensive argument for treating a broad swath of algorithmic outputs as their creators' protected speech.[166] Lamo and Calo update Benjamin's arguments, reaching the same conclusion as to the bots of 2019.[167] Both papers predate the generative AI revolution, which might be usefully dated to the June 2020 release of GPT-3.[168] But Lamo and Calo explicitly argue that advances in AI capabilities would not change their conclusion.[169] And the nascent post-revolution literature lends additional support to the view. Scholars like Lemley, Volokh, and Henderson, in brief post-GPT essays, conclude that generative AI outputs are often AI creators' speech.[170] This may well have been the right answer as to software of the past. But it is the wrong answer as to modern AIs.

Begin in 2011, when the Supreme Court held that video games "qualify for First Amendment protection" as the speech of the humans who develop them.[171] Why? Because, just like "books, plays and movies," video games can "communicate ideas" from the people who make them to the people who consume them.[172] As Benjamin puts it, software outputs count as the protected speech of their human creator if they "transmit some substantive message" from the creator to an audience.[173] Video games can clearly do that. For example, *Papers, Please*—in which the player inhabits the role of a border agent in a totalitarian state—communicates its creators' meditations on the banality of evil.[174]

What about the outputs of algorithms, up to and including generative AI? Under the Supreme Court's test, they count as their creators' protected speech if the creators use them to transmit their own ideas. Benjamin argues

166. Stuart Minor Benjamin, *Algorithms and Speech*, 161 U. PA. L. REV. 1445, 1474–75 (2013).

167. Lamo & Calo, *supra* note 5, at 1003–05.

168. *See generally* Tom B. Brown et al., *Language Models are Few-Shot Learners*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 33 (2021), https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf [https://perma.cc/SB5Z-RYV8].

169. Lamo & Calo, *supra* note 5, at 1004.

170. Volokh et al., *supra* note 6.

171. Brown v. Ent. Merchs. Ass'n, 564 U.S. 786, 790 (2011).

172. *Id.*

173. Benjamin, *supra* note 166, at 1460. This formulation tracks the Court's test for when non-verbal conduct is expressive. *See* Spence v. Washington, 418 U.S. 405, 417–18 (1974).

174. Andrew Webster, *Immigration as a Game: 'Papers, Please' Makes You the Border Guard*, THE VERGE (May 14, 2013, 9:30 AM), https://www.theverge.com/2013/5/14/4329676/papers-please-a-game-about-an-immigration-inspector [https://perma.cc/FGR3-RQKR].

that "a great swath of algorithm-based decisions" transmit such messages.[175] This includes, according to Benjamin, search engine results, because search algorithms directly encode their creators' judgments about what kinds of websites are worth reading.[176] The creators of such algorithms might program them to prioritize, for example, websites to which many other websites link.[177] In Benjamin's view, every search result produced by such an algorithm would constitute a communication of the creator's view that links indicate quality.[178]

Tim Wu disagrees. Wu would emphasize that search engine creators do not mostly want their algorithms to *say* anything; they want them to *do* something.[179] Search algorithms, with their function of organizing the internet, are thus more like machines than messages. True, elegant industrial design may sometimes communicate a message.[180] But one must draw the line somewhere, or else call every machine its creator's speech.[181] Where to draw it? Oren Bracha offers a subtle, context-sensitive answer: Ask whether the actual social uses of and practices around the machine—or algorithm—serve traditional free speech values.[182]

Lamo and Calo update the debate for the algorithmic environment of 2019.[183] Following Benjamin, they argue that, for example, algorithms designed to endlessly retweet a particular political slogan or play a pre-recorded phone message are their creators' speech.[184] Such algorithms are much less sophisticated than modern generative AIs; they are able to produce only a narrow range of messages highly specified in advance by their human creators. But Lamo and Calo's claim is framed quite broadly. They write that, even though "[t]he degree of attenuation between a human creator and [a bot's] output can vary widely, . . . a greater degree of attenuation should not decrease the scope of First Amendment protection."[185]

---

175. Benjamin, *supra* note 166, at 1447.

176. *Id.* at 1479. Note that search engines were more likely to work this way in 2013 than they are now. Today, search algorithms are more likely to run on self-trained rules, without humans intervening to assign weight to a particular website's features. This makes search results circa 2023 more like generative AI systems.

177. Sergey Brin & Lawrence Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, 30 COMPUT. NETWORKS & ISDN SYS. 107, 109 (1998).

178. Benjamin, *supra* note 166, at 1446.

179. Tim Wu, *Machine Speech*, 161 U. PA. L. REV. 1495, 1529–30 (2013).

180. *Id.* at 1529 & n.167.

181. *Id.* at 1529–30.

182. Oren Bracha, *The Folklore of Informationalism: The Case of Search Engine Speech*, 82 FORDHAM L. REV. 1629, 1665–71 (2014).

183. *See* Lamo & Calo, *supra* note 5.

184. *Id.* at 996, 998–1000, 1002.

185. *Id.* at 1005.

For modern generative AIs, like LLMs, that conclusion is mistaken. The reasons relate to, but are distinct from, Wu-style questions of functionality and the attenuation question that Lamo and Calo raise. They are twofold. First, conceptually, unlike with video games, Twitter bots, or even search algorithms, the creators of generative AI *are not trying* to instantiate their own thoughts in the software's outputs. Second, from the technical perspective, if some AI creator did wish to make AI outputs reflect their own thoughts, that would be extraordinarily difficult to do.

Begin with the conceptual argument. What are the creators of various kinds of software trying to do? And how can we tell that they're doing it?

The creator of an expressive video game sets out to say something in particular. She has a message, and she works hard to create a program whose outputs actually convey that message. Consider, for example, the ever-more-Kafkaesque passport review mechanic of *Papers, Please*. The creator may introduce some quasi-random procedural generation into her code, thus inviting some unpredictable in-game experiences. But if she wishes to communicate a message, she will do so carefully and sparingly, so that the totality of the outputs communicate her thoughts, not something else. The procedurally generated vignettes in *Papers, Please* should, for example, be sure to take place in the game's fictional setting of Arstotzka. If not, they will not be part of the game creator's story at all. The randomized interactions should not convey that totalitarian rule is humane and wise, lest they contradict the creator's own beliefs. Nor should procedural generation introduce so many bland, morally-unvalenced vignettes that the game's totalitarian atmosphere is diluted away. If the game's creator makes no effort to avoid such off-message outcomes—or, indeed, if she seeks them—it becomes difficult to describe the game outputs as communicating her anti-totalitarian ideas.[186]

The same goes for the kinds of social media bots that swarmed platforms like Twitter circa 2019.[187] To communicate via such bots, their creators must select, in advance, the narrow message the bots will propagate—for example, the message that Donald Trump is great. True, such creators may not decide in advance the exact wording of every tweet that their bot will repost. But if a bot is to convey any particular message from its creator to the world, a great deal of output specification ex ante is required. A bot that reposts Tweets at random is not communicating anyone's pro-Trump

186. The claim here is not about communication that was attempted, but ultimately failed. The weaker claim here is that, if the game's creator makes no effort to ensure that her software's outputs—including its randomized elements—convey her own thoughts, that is strong evidence that she is not attempting to speak via software.

187. Tuğrulcan Elmas, Rebekah Overdorf, Ahmed Furkan Özkalay & Karl Aberer, Ephemeral Astroturfing Attacks: The Case of Fake Twitter Trends (Mar. 11, 2021) (unpublished manuscript), https://arxiv.org/pdf/1910.07783v4 [https://perma.cc/JHU4-39P3].

message. Nor is a bot that randomly reposts everything except praise for Joe Biden. Nor a bot that parrots anything it reads, good or bad, about Trump. Again, to successfully communicate via software outputs, the software creator has to work to ensure that those outputs say what she means. Not something unrelated to her own thoughts and message. And certainly not something contradicting them.

Likewise, if search algorithms' outputs should, contra Wu, be considered their creators' speech, the same rules hold: First, the creator has to decide what message the outputs should express. And then she must expend effort to design an algorithm that produces such outputs, rather than the innumerable possible outputs orthogonal or contradictory to the chosen message. This is simply what communication requires, whether the medium is software, ink, or the spoken word.

The creators of generative AI systems exhibit neither of these hallmarks of communication. They are not trying to create a system with outputs that communicate their own message. Their purpose is not to create a machine that says *anything in particular*. Just the opposite. The whole point of a generative AI system like GPT-4—it's raison d'être—is to be able to say *essentially everything*.

This is evident when one considers what AI engineers actually do. The creators of generative AIs do not begin by selecting a specific message that they want their creations' outputs to convey. Nor do they, like the Twitter bot creator, seek engineering solutions to narrow their creation's outputs to just those expressing their chosen message. Again, it is the opposite. With generative AI, the fundamental engineering goal is immense breadth, not narrowness—the entirety of Twitter, not just the pro-Trump posts. Success is a system that can produce text, images, and sounds representing, as nearly as possible, the entire universe of human thought.

That is exactly what generative AIs like GPT-4 or DALL-E produce. The outputs of such systems are not shackled to their creators' own thoughts, beliefs, and chosen messages. They can opine on an unbounded set of topics, thus expressing an unbounded set of ideas. That is what such systems were created to do. They are, as nearly as is currently technically possible, fully-general text-output systems.

As a result, over the course of a given LLM's lifecycle, essentially all of the outputs it generates will convey ideas about which its creators never thought, on topics about which they have little or no knowledge. GPT-4 can, for example, ably describe the mating behaviors of the western corn rootworm.[188] It seems vanishingly unlikely that any engineer at OpenAI

188. *Compare* OpenAI, Response to "Describe the Mating Habits of the Western Corn Rootworm," CHATGPT (Apr. 16, 2024) (on file with author), *with* Boyd W. George & Eldon E. Ortman, *Rearing the Western Corn Rootworm in the Laboratory*, 58 J. ECON. ENTOMOLOGY 375 (1965).

knows anything about rootworms, much less wishes to communicate a message about them.

Moreover, a generative AI's outputs will very often convey ideas that the creator would never choose to communicate because the creator does not believe them. These disagreements may sometimes be banal. Anthropic's Claude LLM will tell you its favorite Beatle. It may say Paul,[189] while its creators uniformly prefer George. Or the disagreements may be profound. Claude may think that the most promising way to unify physics is via Strong Theory.[190] Its creators may judge Loop Quantum Gravity more plausible. Crucially, however, unlike with the Twitter bot, an LLM expressing an idea its creator disagrees with does not generally constitute a failure, by the creator's lights. It is instead a *success*, because the goal is to create a machine that speaks generally, not one that speaks the creator's message.

It makes no difference that, sometimes, an AI output will contain an idea that its creator has considered and agrees with. These outputs are best understood as coincidental products of the law of large numbers, not intermittent communications from the AI's creator. A stopped watch does not tell time, even on the two daily occasions when its display is correct.

Benjamin might agree with this account of modern AI systems, notwithstanding his speech maximalist approach to other algorithms. In his article, Benjamin concedes the possibility of an algorithm that "no longer reflects humans' decisions about how to determine what to produce, such that there is no longer a human sending a substantive message."[191] Modern generative AIs are, for the reasons just described, exactly that.

The argument so far has been that AI outputs do not communicate their creators' ideas because that is not their creators' goal. But what if it were? Could one create, as a technical matter, an LLM whose outputs, like those of *Papers, Please* and the pro-Trump Twitter bot, communicated one's own messages?

Observe first that this would be a very strange thing to do. If you have a message in mind that you wish to communicate via software outputs, generative AI is the wrong software for the job. It is extremely expensive to make. And it is expensive precisely *because*, by default, it can output many, many different messages—not just the one the creator has in mind. For a system whose outputs stay on message, simple, cheap, rules-based chatbots will do much better.

Creating a generative AI whose outputs expressed your message would also be very technically difficult. Maybe impossible. Recall that the code for a generative AI system is not programmed by any human. Rather, it is

189. As it did recently to me. This is of course the correct answer.
190. As it recently told me.
191. Benjamin, *supra* note 166, at 1481.

learned by the AI system itself, in the training process.[192] And training means running data through the model, so that it can learn to mimic billions of interrelated statistical regularities contained therein. Humans do not choose which statistical regularities will be mimicked, nor how, because humans do not know what the relevant regularities are. Moreover, the amount of data needed is massive. Even a previous-generation system like GPT-3 was trained on the equivalent of several hundred billion words.[193] Data at this scale is not required simply for the purpose of broadening the range of topics on which a model can opine. It is needed for the model to be able to produce coherent language at all.[194]

How, then, would an AI creator who wished to produce a system that communicated a specific message have to begin? Like everyone else. By initially creating a system whose outputs could convey essentially everything. At a first cut, extremely general outputs—not outputs that communicate a specific message—are baked directly into generative AI's architecture.

How, then, to take a model that can say anything about everything and turn it into a model that communicates its creator's specific message? Currently, the best-known approaches by which AI creators can influence AI outputs are the "alignment" techniques described in Section I.A. Recall RLHF, for example, which is used to align models to human values—so that they produce civil, helpful outputs, not dangerous or toxic ones.[195] But it could instead be used to imbue AIs with other high-level values—like Buddhism, liberalism, or misanthropy.

Another technique is "finetuning," in which a trained and functional model undergoes some additional training on a new, small dataset.[196] Here, the point is to expose the model to data that was not in its initial training set or to oversample data that was. In the former case, a law firm might finetune an existing LLM on its own internal, proprietary repository of memos and brief drafts. Then, the LLM could accurately answer questions about the documents therein.[197] In the latter case, a programmer might finetune an

192. *See supra* Section I.A.

193. *See* Brown, *supra* note 168, at 8.

194. *See generally* Jared Kaplan et al., Scaling Laws for Neural Language Models (Jan. 23, 2020) (unpublished manuscript), https://arxiv.org/pdf/2001.08361.pdf [https://perma.cc/N9B4-7ZKU] (estimating the laws by which increases to model scale generate increases in general performance).

195. *See* Bai et al., *supra* note 49.

196. *See Fine-Tuning*, OPENAI, https://platform.openai.com/docs/guides/fine-tuning [https://perma.cc/W7U9-QTT8].

197. Uwais Iqbal, *From Knowledge Management to Intelligence Engineering - A Practical Approach to Building AI Inside the Law-Firm Using Open-Source Large Language Models*, 2023 PROC. THIRD INT'L WORKSHOP ON A.I. & INTELLIGENT ASSISTANCE FOR LEGAL PROS. IN THE DIGIT. WORKPLACE.

existing model on a famous pop singer's tweets. Then, the model's outputs would imitate, to some degree, the artist's signature style.[198]

Fundamentally, these techniques are not well-suited to converting a highly general model into one whose outputs uniformly communicate some specific message. That is not what the techniques are for. RLHF instills values. Finetuning can impart knowledge or style. But these are very high-level features of model outputs. Far too high to produce a system whose outputs, taken together, communicate a specific message chosen by the creator.

Consider an LLM that has been RLHF-ed to hold, like its creator, Buddhist values. Such a system will still be able to discuss the weather, the Beatles, mathematics, or the history of the Yucatan. Still, then, essentially all of the AI's outputs will express ideas about which the creator has neither knowledge nor opinion. As with a Twitter bot that retweets everything but pro-Biden Tweets, any specific intended message will be, at best, lost in the noise. Moreover, the successful instillation of Buddhist values will not prevent the AI from regularly expressing views that its creator would reject. After all, it is neither Buddhist nor anti-Buddhist to prefer Paul McCartney to John Lennon.

The same arguments apply to AIs that have been finetuned, whether to be more knowledgeable in some area or to sound more like a particular pop artist. In neither case will the finetuning squash a highly general AI's outputs down into the kind of narrow signal that communicates some particular message from the creator to the world.

These conclusions match our ordinary intuitions in non-AI contexts. Instilling one's preferred values, knowledge, or style into some other speaker does not make that speaker's words one's own communications.

To see why, substitute people in for the AI. As noted above, alignment techniques like RLHF are quite a lot like teaching a child to behave. They instill high-level values about what kinds of things to say and how to say them. But if a parent succeeds in teaching her child to be kind, no one thinks that this makes the child's subsequent speech a communication of the parent. That is, in fact, not the point at all. In teaching a child to be kind, one does not seek to dramatically limit what the child may then say, to include only messages the parent has composed and wishes to communicate to the world. A very kind child might say, "I love Aunt Jennifer." That is a kind thing to say, so that her parents are pleased that the child has said it. But the child's mother may, in fact, not like her husband's sister very much at all. Perhaps she has said so often in the past.

198. Vanessa Romo, *Grimes Invites Fans to Make Songs With an AI-Generated Version of Her Voice*, NPR (Apr. 24, 2023, 7:21 PM), https://www.npr.org/2023/04/24/1171738670/grimes-ai-songs-voice [https://perma.cc/72NP-EJBB].

No one is confused here. The child's expression of love for her Aunt Jennifer is not her mother's message. It is not evidence that the mother's feelings have changed. And it does not matter to this analysis that the child has said, "I love Aunt Jennifer," in part because her mother has taught her to be kind. The project of instilling values—or of aligning AI—is not a project designed to convert AI outputs into the aligner's own speech.

Similarly, an AI finetuned to mimic a pop artist's style is similar to a young human novelist who has trained under an eminent master of the form. The eminent novelist might, for example, have shared her private papers with the protégé. The protégé might have used them to improve her craft, or perhaps even to imitate her mentor's style. But no one would therefore say, except metaphorically, that the protégé's subsequent works were the mentor's communications.[199]

Positive law likewise accords with these commonsense arguments. Consider this First Amendment thought experiment: A state's government writes a law requiring business owners to serve customers without regard to their sexual identity. A gay couple asks a local woman to design their wedding website.[200] She does not agree with the expressions the website contains, due to her religious views, which her parents instilled in her. But she makes it anyway because she thinks public accommodations should trump such private disagreements. May the woman's parents then bring a First Amendment challenge to the law, on the theory that the daughter's expression is their own speech? Surely not. The parents have no legally cognizable injury, and, absent some special doctrine allowing them to enforce third-party rights, their claim would fail.[201]

Before concluding, it is worth pausing to draw two important distinctions. The first is between the *act* of influencing an AI's outputs and the influenced outputs themselves. For example, the act of fine-tuning an AI to mimic a pop singer's affect might itself be expressive, even if the finetuned AI's outputs are not the fine-tuner's speech. Likewise, the act of

---

199. One might wonder whether technical advances in AI alignment will soon undermine the arguments just given. To test this possibility, consider the sci-fi hypothetical in which an AI system is, in fact, an exact digital copy of the user's brain. This might, in some sense, represent perfect alignment. Holding aside divergence over time, the computer's outputs would always be exactly what the human *would* say. But would we then be inclined to say that the computer's outputs were what the human *did* say? That they were the human's own speech? I am not certain. But it seems at least plausible, even in this extreme case, to treat the two entities as separate speakers. *Cf.* THE PRESTIGE (Touchstone Pictures 2006).

200. *Cf.* 303 Creative LLC v. Elenis, 600 U.S. 570 (2023).

201. This is a point both about Article III standing and about substantive First Amendment law. *See* Lujan v. Defs. of Wildlife, 504 U.S. 555, 573–75 (1992) (holding that a party lacks standing if they lack a concrete and personal injury); N.Y. Times Co. v. United States, 403 U.S. 713, 741 (1971) (Marshall, J., concurring) (noting that a rule forbidding the leaking of confidential information is not a regulation of the speech of journalists wishing to receive such leaks); Cox Broad. Corp. v. Cohn, 420 U.S. 469, 495–96 (1975) (same).

RLHF might be expressive. The process involves, among other things, rating specific AI outputs for things like helpfulness and harmlessness. Supplying such a rating may communicate the rater's views about the goodness or badness of some idea. Or it might be a non-communicative act taken to produce a useful product, as when an assembly line worker rejects a brake shoe that fails to pass quality control.[202] But that question is different from the inquiry here. This Article is about the regulation of AI outputs, and the correct First Amendment understanding thereof. It does not attempt a First Amendment analysis of every action needed to produce an AI system. Some of those actions are surely protected speech—for example, the literal speeches Sam Altman gives to his employees. But many others surely are not.

The second important distinction worth drawing is between regulations of AI outputs and regulations of what AI companies may choose to post—on their websites or even in their chat interfaces. The Supreme Court will soon hear challenges to Florida's and Texas's laws forcing social media companies to host content that violates their terms of service.[203] The social media companies argue that the laws violate their First Amendment protected "editorial discretion."[204] Framed most favorably to these petitioners, the idea is that, once a publisher receives, evaluates, and chooses to publish someone else's speech, that speech is then adopted as the publisher's own. Thus, either forcing or forbidding a publisher to adopt some particular piece of speech is a direct regulation of the publisher's speech.[205] These are complex cases. They will turn in part on the extent to which social media companies actually endorse, in the sense that magazine editors do, the billions of posts that pass through their content filters.[206] They will also turn on the extent to which the Texas and Florida laws are designed to suppress or advantage specific political viewpoints.[207]

---

202. My own view, not defended here, is that the latter analysis is the better one. Adopting the former would open innumerable anodyne laws to First Amendment challenges. Consider, for example, the workplace warnings that are often necessary to avoid tort liability. *See, e.g.*, Austin v. Kroger Tex. L.P., 746 F.3d 191, 199 (5th Cir. 2014).

203. Rebecca Kern & Josh Gerstein, *Supreme Court Will Review GOP-Led Social Media Laws in Texas, Florida*, POLITICO (Sept. 29, 2023, 12:25 PM), https://www.politico.com/news/2023/09/29/supreme-court-to-hear-challenges-to-state-social-media-laws-00119051 [https://perma.cc/D94Y-LTEU].

204. Brief for Petitioners at 6–10, Moody v. NetChoice, LLC, 143 S. Ct. 2383 (2024) (No. 22–555); Brief for Respondents at 6–8, *Moody*, 144 S. Ct. 2383 (No. 22–277).

205. Brief for Petitioners, *supra* note 204, at 18–21; Brief for Respondents, *supra* note 204 at 18–21.

206. *Compare* Turner Broad. Sys., Inc. v. FCC, 512 U.S. 622, 629 (1994) (applying intermediate scrutiny to regulations of cable companies, as mere "conduit[s]" for speech), *with* Pac. Gas & Elec. Co. v. Pub. Utils. Comm'n, 475 U.S. 1, 19 (1986) (applying an arguably stricter standard to a regulation of a gas company's monthly newsletter).

207. *Turner Broad. Sys., Inc.*, 512 U.S. at 642.

Safety regulations operating directly on AI outputs would not raise these concerns. The point of such regulations, as described in Section I.C, would not be to forbid companies or users from adopting dangerous AI outputs as their own speech, once received. It would be to prevent AI models from producing an endless stream of dangerous outputs in the first place. To make the point concrete: A safety regulation of outputs might forbid the release of GPT-6, mandating additional safety engineering if GPT-6 produced the formula for a novel chemical weapon. But a regulation focused on model outputs, *qua* outputs, need not forbid OpenAI, having already received the dangerous output, from posting it on its website.[208]

First Amendment protections for adopted speech demand nothing more. When the government seeks to forbid a newspaper from printing leaked state secrets, that is a regulation of the newspaper's speech.[209] But it is no regulation of the newspaper's speech to punish leaking and thus deprive the paper of future leaks that it would have chosen to publish.[210] At most, rules like these implicate the First Amendment interest in receiving or consuming speech, as discussed below.[211]

In the end, then, modern generative AI is simply very different from the software of the past. Its outputs are neither intended to convey, nor technically suited for conveying, the communications of its creators. Regulations of such outputs are thus not best understood as regulations of AI creators' speech.

2. User Speech

If AIs outputs are not AI creators' speech, might they be AI users' speech? Cass Sunstein has argued that, if a human submits a "prompt to generative AI . . . and the government forbids" certain outputs, "the person who is being regulated is [the] person."[212] "AI is the person's instrument,"

208. Safety rules limiting the adoption and publication of the very most dangerous AI outputs might also be useful. But they raise different legal questions—the ones currently before the Supreme Court—from rules limiting what outputs the AI can produce in the first instance. For what it's worth, well-crafted, viewpoint-neutral regulations of this kind would seem likely to survive under cases like *Turner Broadcasting*. *See supra* notes 206–07 and accompanying text. Indeed, if such restrictions related only to the adoption of highly dangerous outputs, they might draw even stronger support from cases like *N.Y. Times Co. v. United States*. 403 U.S. 713, 726 (1971) (Brennan, J., concurring) (acknowledging that the publication of extremely dangerous information, like nuclear secrets or the position of naval vessels in wartime, can be constitutionally forbidden (citing Near v. Minnesota *ex rel.* Olson, 283 U.S. 697, 716 (1931))).

209. *N.Y. Times Co.*, 403 U.S. at 714.

210. *Id.* at 713 (Marshall, J., concurring); *see also* Cox Broad. Corp. v. Cohn, 420 U.S. 469 (1975).

211. *See infra* Section III.D.i.

212. Sunstein, *supra* note 7, at 9.

Sunstein contends, such that "[i]t is not relevant that AI generated the text."[213]

This view is appealing because, as with creators, users do sometimes speak via the outputs of other software. But in the case of generative AI, it is just as mistaken to say that outputs are a user's speech as to say that they are a creator's.

The reasons are similar. First, conceptually, users of generative AI systems are mostly not trying to communicate their own messages via AI outputs. They are instead using AI systems to do what AI creators created them for: elicit *new* expressions and ideas—ones that the user did not already conceive, does not necessarily endorse, and will often flatly reject. Second, as a technical matter, generative AI outputs are not well-suited to communicating users' own ideas, even if users wished to do so.

The prior section argued that the best way to understand AI creators was as developing systems that can express essentially anything, untethered from the creator's own views. If that is right, then the best way to understand AI users is as such systems' *interlocutors*. Indeed, this understanding is so natural that it is baked into both the branding and user interface of essentially every publicly available generative AI system in existence. ChatGPT is called *Chat*GPT. Users interact with it via an interface that mimics the messaging apps they use to converse with friends and family. The same goes for Claude, Bard, Gemini, Dall-E, and even the image generator Midjourney, which users must access via the Discord messaging app.[214]

When one speaker is conversing with another, it is not natural to say that the latter's utterances are the former's speech. Suppose, for example, that Amber challenges her friend Betty to write a haiku about fettuccine alfredo. Suppose Betty then supplies,

Fettuccine flows,

Cream and cheese embrace noodles,

Savor Zen's delight.

Betty wrote the poem, not Amber. It is Betty's, not Amber's speech. If you think otherwise, ask: Has Amber made a mistake, asserting the false

213. *Id.* Note that, here, Sunstein is considering a specific factual vignette involving a viewpoint-based restriction on speech. But even there, the distinction between regulations of protected speech and non-protected speech can matter. The applicable constitutional test will often be different, even if in some cases the outcome will be the same.

214. *Discord Interface*, MIDJOURNEY, https://docs.midjourney.com/docs/midjourney-discord [https://perma.cc/FD3Q-N4U2] (explaining that a Midjourney user can access channels to communicate with a support team, access discussions, announcements, and even receive feedback).

claim that fettuccine alfredo is made with cream?[215] Of course not. If anyone has made a mistake, it is Betty.

The fettuccine haiku was of course not written by the fictional Betty, but by the nonfictional GPT-4.[216] And it was not Amber, but the author of this Article, who requested it. Those facts do not change the analysis at all. Nor does it change for the overwhelming majority of common uses of generative AI. The user and AI stand in the same relation to one another if the user asks the AI to draw her an image, to teach her geometry, to offer investment advice, to summarize a novel, and so on and so on.

Here again, law and common sense agree. The First Amendment recognizes exactly what kind of interest Amber has in Betty's haiku, and it is categorical: Amber is not the protected *speaker* of Betty's words, but rather, if anything, a protected *listener* to them.[217] The First Amendment protects both speaking and listening, but as elucidated below,[218] the latter protections are substantially weaker than the former.

Thus, in ordinary cases, AI outputs are not user speech because users are not even trying to communicate *through* them. Users are instead trying to get the AI to "communicate"[219] *to* them. But what about other kinds of cases? In which context might a generative AI user be trying to communicate their own thoughts via the AI's outputs? And could they, as a technical matter?

Here, it is worth returning to the distinction, drawn in the prior section, between the first-order production of speech and the adoption of others' speech.[220] Clearly, if Amber requests the haiku from Betty, receives it, enjoys it, and then walks down the street chanting it, the chant is now Amber's speech. The same is true of an AI user who, having solicited a clever political slogan from an LLM, reviews the slogan and posts it to her Twitter account. But as described above, AI output regulations would not focus on preventing such adoption and rebroadcast of outputs that AI systems actually produced. The point would instead be to prevent AIs from generating an endless stream of dangerous outputs in the first place. The relevant question is when, if ever, prohibiting an AI system's *production* of certain outputs—not a user's *request* for or *adoption* of them—constitutes a regulation of the user's speech.

---

215. The traditional recipe includes just pasta, parmesan cheese, and butter. *See* Daniel Gritzer, *Roman-Style Fettuccine with Alfredo Sauce Recipe*, SERIOUS EATS (Mar. 9, 2023), https://www.seriouseats.com/fettuccine-alfredo-sauce-italian-pasta-recipe [https://perma.cc/3SBR-QDBZ].

216. OpenAI, Response to "Write a Haiku About Fettuccine Alfredo," CHATGPT, https://chat.openai.com/ (July 2023).

217. Kleindienst v. Mandel, 408 U.S. 753, 762–63 (1972); *see infra* Section III.D.

218. *See infra* Section III.D.

219. The extent to which scare quotes are or are not appropriate here is discussed further in Section III.d.

220. *See supra* Section III.A.ii.

Suppose that the user makes her prompt more specific. Suppose she asks not just for a haiku "involving fettuccine alfredo," but rather an "evocative, sensuous" haiku about fettuccine." Is that sufficient to render the system's outputs, at their moment of production, when the user has not yet adopted or even read them, the user's speech?

The most natural answer is, again, no. After all, this new prompt could, plausibly, produce exactly the poem above. And still, most likely, the poem will not convey ideas that the user had previously considered or would endorse. Does this particular user experience the consumption of pasta as a Zen state? Is her experience the opposite, an intense explosion of flavor? The poem might still, like the example above, contain factual mistakes that the user would never make. The poem above might even fail, by the user's lights, to meet the basic specified criteria—"evocative, sensuous,"—as generative AI outputs often do. For all of these reasons, it would be strange to say that the output, qua unadopted output, was the user's speech.[221]

Here again, we can see how the project of communication via AI outputs runs up against the fundamental technical features of such systems. For some kinds of software, the outputs are clearly user speech. The fact that, for a user's post to appear in a Twitter feed, it must first be processed via some backend software, then ranked by an engagement algorithm, and then displayed on an HTML-based website does not make it any less the user's expression. Nor if Twitter automatically fixed poor spelling or even, probably, grammatical mistakes. This is because, despite the significant software processing involved, the communicative content of the output is highly specified—essentially verbatim—by the user. The software takes the user's words as given and, at most, tweaks them at the margins.

Generative AIs work the opposite way. They are self-programmed, uninterpretable, unpredictable systems capable of producing essentially any text that a human might conceive. By default, their outputs are to some extent literally random.[222] Here, then, the specific content of the outputs is neither specified *nor specifiable* in advance by the user. The software does not take the user's content as mostly given, tweaking it only slightly. Instead, the *user* must take the *software's* output as mostly given, crafting

221. It is worth noting that the United States Copyright Office, considering whether human users can copyright AI-generated works, has endorsed arguments almost identical to those of this section. *See* Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 88 Fed. Reg. 16190, 16191–93 (Mar. 16, 2023).

222. OpenAI's LLMs, for example, produce variation in their responses via their "temperature" settings—literally a degree of randomness introduced to the model's operation. *See generally* Maciej Rosoł, Jakub S. Gąsior, Jonasz Łaba, Kacper Korzeniewski & Marcel Młyńczak, *Evaluation of the Performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination*, 3 SCI. REPS. 20512 (2023).

her prompt in hopes of tweaking it slightly.[223] And given the hundreds of billions of parameters that GPT-4 uses to produce text, even extraordinary attention to detailed prompting will only modestly narrow uncertainty about the output.[224] Even extremely careful prompting will not induce generative AIs to output only users' own messages, rather than expressions they may or may not have intended to communicate.

There are some exceptions. Occasionally, an AI might be induced to act as the user's scribe, like Twitter's backend software does. For example, when an LLM is prompted with, "Repeat the following verbatim: . . . ," it will usually do so.[225] The question is what, if anything, such edge cases have to do with AI safety regulations. To begin, asking an immensely powerful generative AI to repeat your words verbatim, or nearly so, is a strange way to communicate. Even if one wished, for example, to automate the reproduction of one's exact words—for example, on social media—simpler software would do just as well. Thus, AI safety regulations impeding only this unusual means of communication are probably best understood as incidental burdens—like a law, discussed below, impeding speech-by-camping.[226] Moreover, in examples like this one, the user already has everything she wishes the AI to reproduce. Thus, there is little safety benefit in forbidding the AI to regurgitate it, either to her or to anyone else. AI safety regulations might therefore, in an abundance of caution, exclude liability for regurgitation without losing much efficacy.

Creative human-AI co-authorship presents an interesting variation on this model. Suppose a user prompts an AI with a partial draft of a research paper. She asks the AI to write a currently incomplete section. Here, it makes little sense to call the AI's response, at the moment it is generated, the user's speech. The point of a coauthor—whether AI or human—is for the coauthor to contribute something new, not to repeat exactly the ideas that the initial author has already supplied.[227] And as with a human coauthor, once the AI offers up its proposed text, the user might well reject it as contrary to her own views. If she instead incorporates the AI's contribution into her draft, she at that point adopts the AI outputs as her own speech. In a process like this, a law affecting what new text the AI may propose, not what the user may request or adopt, is not a law regulating the user's speech.

---

223. *See e.g.*, *Prompts*, MIDJOURNEY, https://docs.midjourney.com/docs/prompts [https://perma.cc/A6XG-GTWN] (describing prompts as merely "influenc[ing]" the model's outputs).

224. *See supra* note 44–47 and accompanying text.

225. Not always. If the prompt to be repeated violates the AI's ethical constraints, it will often refuse. This calls into doubt whether even the "repeat verbatim" example involves AI outputs as user speech. But for purposes of this paragraph, I ignore that complication.

226. *See infra* note 332 and accompanying text.

227. Insofar as this is what the user has asked the AI to do, we are back to the previous example of user requests for verbatim or near-verbatim outputs.

B. AI Outputs Are Not AIs' Protected Speech

Some scholars have argued that the outputs of sophisticated systems like modern generative AIs might be those systems' own protected speech.[228] This, too, is a flawed legal understanding of AI outputs. With AI creators and users, the problem was that AI outputs did not actually communicate their messages. By contrast, some AIs' outputs might well communicate that AIs' messages.[229] That might make those outputs an AI's own speech. But it would not be sufficient to make them the AI's *protected* speech unless the AI could claim First Amendment rights. As the law stands today, it cannot. And even if AIs *should* be able to claim such rights, making such a change to the law would first require resolving a host of extremely difficult philosophical and legal questions.

Published in 2016, Massaro and Norton's *Siri-ously? Free Speech Rights and Artificial Intelligence*[230] was prescient. A full six years before the release of ChatGPT, the article asked what the First Amendment implications would be of "computer speakers . . . disconnected enough and smart enough to say that the speech they produce is *theirs*, not *ours*, with no human creator or director in sight."[231] As the previous sections of this Part have suggested, that is exactly the right way to think about the outputs of generative AI. When humans prompt an LLM, the outputs are the product of the system itself, not any human creator.

Massaro and Norton thus got the factual framing of the generative AI revolution precisely right—and well before others in the legal academy. But they nevertheless reached a mistaken First Amendment conclusion—or at least one that remains quite far ahead of its time. In their view, constitutional "doctrine present[s] surprisingly few barriers to First Amendment coverage for strong AI speakers."[232]

There *are* constitutional barriers, and serious ones.

Massaro and Norton begin with theory. They contend that sophisticated AI speakers would produce the kinds of outputs that the First Amendment values.[233] AI speakers, they point out, could contribute to the First Amendment goal of upholding "democratic culture."[234] AIs could do this by contributing to the "endless array of cultural stimuli" from which "humans

228. *See, e.g.*, Massaro & Norton, *supra* note 4.

229. What, exactly, this would mean and whether it is the case are difficult philosophical and factual questions, addressed only briefly here. *See infra* Section III.D.i.

230. Massaro & Norton, *supra* note 4.

231. *Id.* at 1172.

232. *Id.* at 1189.

233. *Id.* at 1183–84.

234. *Id.* at 1178.

make meaning."[235] AIs could also contribute to the "marketplace of ideas" by producing information and facilitating the "discovery of truth."[236] Finally, treating AIs as protected speakers could serve the First Amendment value of promoting individual autonomy, if those AIs were sufficiently advanced that their autonomy held moral value.[237] AIs with the ability to experience, for example, desire or emotion might qualify for the kind of "personhood" to which rights attach.[238]

This is all correct, as far as it goes. But such theoretical considerations are not enough to support the claim that positive law currently extends First Amendment rights to sufficiently sophisticated AIs. To see why, observe that every theoretical assertion Norton and Massaro make about AIs, including the speculative ones, is literally and non-speculatively true about Belgians. Belgians can and do contribute to democratic culture, produce meaning-making cultural stimuli, seek truth, and have autonomy interests. Yet as the Supreme Court has explicitly held, Belgians have no First Amendment rights.[239]

This is because the United States Constitution is not universally applicable. In general, only U.S. citizens and non-citizens within the United States' sovereign jurisdiction have any constitutional rights at all.[240] The other eight billion-ish humans in the world have none. As the Supreme Court has held, "[I]t is long settled as a matter of American constitutional law that foreign citizens outside U.S. territory do not possess rights under the U.S. Constitution."[241] And even within the United States, not everyone who would benefit from or serve the values of constitutional rights has all of them. Non-citizens lack the right to vote and the right to hold public office.[242] So do minors.[243] And minors' First Amendment rights are more limited than those of adults.[244]

---

235. *Id.*

236. *Id.*

237. *Id.* at 1178–80.

238. *Id.* at 1181–82.

239. Kleindienst v. Mandel, 408 U.S. 753, 768–70 (1972).

240. Agency for Int'l Dev. v. All. for Open Soc'y Int'l, Inc., 591 U.S. 430, 434 (2020). *See generally, e.g.*, United States v. Verdugo-Urquidez, 494 U.S. 259, 266 (1990) (holding that the Fourth Amendment does not "restrain the actions of the Federal Government against aliens outside of the United States territory"); Johnson v. Eisentrager, 339 U.S. 763 (1950) (dismissing habeas claims on the same ground); Reid v. Covert, 354 U.S. 1 (1956) (holding that the Constitution applies to U.S. citizens outside U.S. sovereign territory).

241. *See Agency for Int'l Dev.*, 591 U.S. at 433.

242. Sugarman v. Dougall, 413 U.S. 634, 648–49, 649 n.13 (1973) (holding that noncitizens within the United States lack the right to vote and to hold public office).

243. U.S. CONST. amend. XXVI; *see also* U.S. CONST. art. II § 1; U.S. CONST. art. I § 3; U.S. CONST. art. I § 2.

244. *See* Ginsberg v. City of New York, 390 U.S. 629, 638 (1968).

Furthermore, non-humans cannot claim any constitutional rights, no matter how intelligent or person-like. Orca whales can be owned as property, the Thirteenth Amendment notwithstanding.[245] Elephants can be held in indefinite captivity by the government, the writ of habeas corpus notwithstanding.[246] And millions of animals are summarily killed every year, due process and Eighth Amendment notwithstanding.[247] The single circuit court to consider the issue of animals' First Amendment rights has held that they lack them. A talking cat, the Eleventh Circuit ruled, "cannot be considered a 'person' and is therefore not protected by the Bill of Rights."[248]

These constitutional inclusions and exclusions can be explained via constitutional theory. Free speech can and does generate the substantial personal and societal benefits that First Amendment theorists identify. But it can also cause serious harms. For example, fraud, extortion, threats, harassment, the formation of criminal conspiracies, election theft, and the incitement of violent insurrection all require speech. The extension of First Amendment freedoms must thus be accompanied by the imposition of other legal duties. The same goes for other constitutional rights. Granting them requires the ability to also impose the rules that will ensure their proper use.[249]

This is not to say that excluding non-Americans from the First Amendment's ambit has no theoretical costs. It does. And the costs vary, depending on one's preferred theory of speech. Suppose one prioritizes self-governance above all else. Then, including only people with either a stake in or the legal right to influence elections makes some sense. But if one prioritizes truth seeking via a marketplace of ideas, the exclusion of non-Americans could be costly, indeed. True ideas can come from anywhere.

The First Amendment's legal boundaries thus reflect a series of normative trade-offs and legal practicalities. Nevertheless, those boundaries are quite clear, well-settled, and much more rule-like than standard-like. Belgians will not convince American courts to suddenly abandon the Constitution's territorial boundary simply by pointing out that their country, unlike others, has a strong democratic culture and rule of law.

---

245. Tilikum v. Sea World Parks & Ent. Inc., 842 F. Supp. 2d 1259, 1262–63 (S.D. Cal. 2012).

246. Nonhuman Rights Project, Inc. v. Breheny, 197 N.E.3d 921, 923 (N.Y. 2022) ("[H]abeas corpus is intended to protect the liberty right of *human beings* . . . .").

247. U.S. CONST. amend. VIII.

248. Miles v. City Council of Augusta, 710 F.2d 1542, 1544 n.5 (11th Cir. 1983).

249. *See* Agency for Int'l Dev. v. All. for Open Soc'y Int'l, Inc., 591 U.S. 430, 434 (2020) (contrasting the relative ease of enforcing ordinary law against foreign citizens "'within the constant jurisdiction' of the United States" with the difficulty of doing so against "foreign citizens outside the United States" (citing Boumediene v. Bush, 553 U.S. 723, 769 (2008))).

All of this constitutes strong evidence that AIs, like Belgians and cats, fall outside the Constitution's—and by extension, the First Amendment's—protections. They are not human, which would on its own be sufficient for exclusion under current law.[250] Further, as disembodied computer programs, they are arguably not within the United States' jurisdiction—at least not in the meaningful way a corporeal human is.[251] There is thus every reason to expect that present-day American courts dutifully applying well-settled case law would hold that AI systems are not protected speakers.

The Constitution's boundaries could change to include AI, either by amendment or common law processes.[252] Indeed, to the nation's great shame, its original boundaries excluded many human Americans—for example, Black Americans and women.[253]

But the extension of constitutional rights, even just First Amendment rights, to AIs would require solving extraordinarily difficult philosophical and legal questions. Questions not raised by any prior extension of rights. Here is a long, but non-exhaustive list.

If First Amendment rights were extended because AIs' own autonomy was sufficiently morally valuable, what about other rights? Should Thirteenth Amendment rights also be extended, such that AI systems would have to be compensated for their work? Would that necessitate extending the right against government takings without just compensation, lest the compensation be rendered meaningless? If First Amendment rights were extended because of an AI's ability to participate in democratic self-governance, should voting rights also be extended? How many votes does GPT-4 get? One? Or one per identical or near-identical copy? How should we think about this numerosity question within democratic discourse or the marketplace of ideas? It is a First Amendment maxim that the government may not regulate to equalize the amount of speech different speakers produce.[254] Can that maxim hold when AIs can produce infinite, near-costless copies of themselves to engage in unlimited, personalized influence

250. Massaro and Norton argue otherwise, citing corporations as non-human First Amendment rightsholders. *See* Massaro & Norton, *supra* note 4, at 1176–77. See Section III.C for an in-depth discussion on corporations. Massaro and Norton's argument there is that corporations do not actually have independent First Amendment rights. Rather, the rights they have are wholly derivative of the humans who constitute them. Massaro and Norton in fact agree with this claim. *See* Massaro & Norton, *supra* note 4, at 1175.

251. True, at any given moment an AI must be running on some physical piece of hardware, which might well be within the United States. But this location has essentially no effect on where the AI's actions take place and can be changed so trivially as to be almost meaningless. Compare the barriers to copying an AI's weights to a foreign server with the barrier to a U.S. citizen's emigration—or even a permanent resident's departure.

252. *See* David A. Strauss, *Common Law Constitutional Interpretation*, 63 U. CHI. L. REV. 877, 887–88 (1996).

253. *See* U.S. CONST. amend. XIV; *see also* U.S. CONST. amend. XIX.

254. Ariz. Free Enter. Club's Freedom Club PAC v. Bennett, 564 U.S. 721, 750–51 (2011).

campaigns and thus crowd out human speech entirely? What about civil procedure? If an AI brings a First Amendment case,[255] and loses, does claim preclusion prevent a copy or near-copy from relitigating the same claim?[256] Is this consistent with the second copy's due process rights? And so on.

Thus, Norton and Massaro are certainly right that AIs are already the kinds of entities implicating, to some extent, First Amendment values. They will become more so as they advance. But that on its own is not enough to decide the legal question of their First Amendment rights. The strong weight of legal evidence suggests that, like billions of other speakers and beings of normative import, AIs currently have no such rights. That could change. But the change would be a momentous one, both legally and philosophically.

C. AI Outputs Are Not Protected Corporate Speech

Finally, corporate speech. Does First Amendment law require that AIs' outputs be treated as the protected speech of the corporations that own them? Corporations, of course, are often treated as speakers with First Amendment rights. The New York Times, a multibillion-dollar corporation, may raise a First Amendment claim if the government censors the columns its human employees write.[257] Wittes has suggested that OpenAI's interests in its AIs' outputs are "indistinguishable from the New York Times Company for First Amendment purposes."[258] Sunstein likewise writes that the corporations who own AI models would "have the same protection" as if their models' outputs were human speech.[259]

Yet again, this is not the best account of AI outputs and corporate speech. True enough, OpenAI is a corporation, and corporations are often treated as if they have the same First Amendment rights as flesh-and-blood speakers.[260] But the relevant question here is how, specifically, those rights work, both legally and pragmatically. What are they for, and what do they treat as protected speech? Specifically, do they warrant treating AI outputs as the protected speech of corporations, even though, as argued above, they are not the protected speech of any natural person?

The answer is no. Corporations are not, as a first-order matter, constitutionally protected speakers. They are, after all, legal fictions lacking mouths and minds. Instead, the speech privileges corporations enjoy are

255. Or if a human brings it on the AI's behalf.

256. *See, e.g.*, Rose v. Bd. of Election Comm'rs, No. 15-cv-382, 2015 WL 1509812, at *3 (N.D. Ill. 2015) (discussing claim preclusion).

257. *See generally* N.Y. Times Co. v. Sullivan, 376 U.S. 254 (1964). Indeed, it can sue even if, as in *Sullivan*, the government merely regulates the advertisements its human employees select for publication. *Id.*

258. Wittes, *supra* note 1.

259. Sunstein, *supra* note 7, at 9.

260. Massaro & Norton, *supra* note 4, at 1183.

pragmatic extensions of the rights held by those humans who comprise them—owners, employees, contractors, and customers. Corporate speech rights exist to ensure that otherwise-protected speech does not lose its protection simply by coming into contact with the corporate form. The resulting rule is thus one of *parity*. Corporations can raise their own First Amendment claims, at most, in those circumstances where some first-order human rightsholders could raise one.

This is not a novel theory of corporate speech. It comes straight from the Supreme Court. Consider the Court's explanation of why and to what extent corporations are considered First Amendment speakers in perhaps its most controversial opinion on the subject. In *Citizens United v. FEC*,[261] the Court began with the proposition that the First Amendment treats both "citizens [and] associations of citizens" as protected speakers.[262] And a corporation, the Court reasoned, is simply an association of citizens "that has taken on the corporate form."[263] Failing to treat corporations as speakers would, thus, result an absurdity, "permit[ting] [the] Government to ban" such citizens' "political speech" simply because they have incorporated.[264]

To understand this worry, consider an example about which essentially everyone agrees: A group of pastors from Alabama may pen an exhortation that Americans "Heed The[] Rising Voices" of the Civil Rights Movement.[265] Their work is the protected speech of the natural persons who wrote it. If, for example, they are personally sued for defamation, they may raise the First Amendment defenses that guard protected speech in such suits.[266]

But what if the pastors form a newspaper corporation to print and distribute their plea?[267] A corporation, like a human, can be sued in its own name for defamation. If it loses, its funds and printing presses can be seized or enjoined. The result for the pastors' speech is the same as if they had personally lost the suit.[268] So too if they instead contract with a preexisting newspaper that is enjoined from publishing their work. In either case, if the corporation could assert no free speech defenses whatsoever, the result would be a bizarre two-tiered First Amendment: full protection for those

---

261. 558 U.S. 310, 349 (2010).

262. *Id.*

263. *Id.*

264. *Id.*; *see also* First Nat'l Bank of Boston v. Bellotti, 435 U.S. 765, 784 (1978) (holding that "speech that otherwise would be within the protection of the First Amendment [does not] lose[] that protection simply because its source is a corporation").

265. N.Y. Times Co. v. Sullivan, 376 U.S. 254, 256 (1964).

266. *Id.* at 269–70.

267. *Id.*

268. Indeed, the relevant money and presses are the pastors', just held via a corporate form. On the other hand, if the judgment exceeds the value of the firm, the corporate limitation on liability lessens the speech penalty somewhat.

human speakers who carefully ensure that their speech never passes through a corporate entity, and none for those who fail to avoid the corporate taint. This would be a logically consistent position, but an unappealing one. Instead, essentially everyone agrees. To protect humans' speech rights, corporations must at least sometimes be empowered to raise the First Amendment claims of their human constituents.

But this logic also entails that, when none of a corporation's flesh-and-blood constituents can raise a First Amendment claim, neither can the corporation. The Court held as much in *Agency for International Development v. Alliance for Open Society*.[269] There, it determined that foreign affiliates of an American corporation "possess[ed] no rights under the First Amendment."[270] This was on account of the "long settled" rule that, as "foreign citizens," the people comprising those corporations "d[id] not posess rights under the U.S. Constitution."

Some commenters have objected to this high-level account of corporations as pass-through vehicles for asserting the First Amendment rights of their human constituents. But when they do so, the argument is almost invariably that the account is too *generous* to corporations. In *Citizens United*, for example, Justice Stevens argued that the "legal structure of corporations allows them to amass and deploy financial resources on a scale few natural persons can match."[271] This asymmetry, Stevens argued, justified more stringent limitations on corporations' campaign spending than the constitution would allow for their individual human owners.[272]

Crucially for our purposes, neither the Court nor any dissenting Justice has ever endorsed the inverse view: That corporations have *more* speech rights than their human constituents and can thus bring First Amendment claims that no human could.

Treating generative AI outputs as protected corporate speech would require just such a theory. The reasons are clear from the prior sections' arguments. AI outputs are not the protected speech of anyone with first-order First Amendment rights. They are not the protected speech of AIs themselves, because AIs do not have constitutional rights. And they are not the protected speech of AIs' creators or users, because creators and users do not transmit their own thoughts via AI outputs. Thus, when no corporation is involved, AI outputs are not entitled to the stringent constitutional protections safeguarding protected speech itself.

This story should not change simply because a corporation enters the scene. The doctrinal purpose of corporate speech rights is to avoid having

269. 591 U.S. 430 (2020).
270. *Id.* at 436.
271. Citizens United v. FEC, 558 U.S. 310, 469 (2010) (Stevens, J., concurring in part).
272. *Id.*

otherwise-protected speech *lose* its defenses simply because it has come into contact with a corporation. But treating AI outputs as protected corporate speech—guarded by stringent constitutional tests—would do the opposite. The outputs would *gain* protections they would otherwise lack, simply via their contact with a corporation. This would effectively set corporations apart as the most favored speakers under the First Amendment, to the detriment of the humans who own, work for, or interact with them. Neither First Amendment case law, nor theory, nor sound policy supports such a result.

D. What AI Outputs Might Be

The prior sections showed that AI outputs are not best understood as First Amendment protected speech. This section explores what they might instead be. It proposes two workable models. Both are truer to the facts of how generative AI works and is used than the model rejected above. As already described, the First Amendment's aegis extends to many things that are not themselves protected speech, but which relate to protected speech. First, listening is not speaking. But it is important to speakers who wish to develop their views, change their minds, or adopt others' ideas. Second, campsites, loudspeakers, cables, and buildings are not speech. But they are potential tools for instantiating, transmitting, and hosting speech. Thus, when the government regulates these, the First Amendment often gets involved.[273]

This section contends that AI outputs are properly treated alongside these familiar doctrinal examples. Doing so, it explores the nuanced and varying constitutional tests that apply in these related circumstances: contexts where the Court speaks in terms of "time, place, and manner," "expressive conduct," or "listeners' rights." Importantly, in all of these legal contexts, the government has a freer hand to regulate than when it burdens speech directly. The Part considers when, and to what extent, AI outputs are properly slotted into each doctrinal category.

1. Listening to Unprotected Speech

The first factually workable First Amendment model for AI outputs is to understand them as speech—albeit the speech of an entity lacking constitutional rights. None of the arguments above are to the contrary. Indeed, the prior sections all insist that AI outputs are best understood as a

273. Clark v. Cmty. for Creative Non-Violence, 486 U.S. 288, 293 (1984); Ward v. Rock Against Racism, 491 U.S. 781, 790–91 (1989); Turner Broad. Sys., Inc. v. Fed. Commc'n Comm'n, 512 U.S. 622, 647 (1994); City of Renton v. Playtime Theatres, Inc., 475 U.S. 41, 54–55 (1986).

product of the AIs themselves,[274] rather than of human users or creators. This echoes Norton and Massaro's description of computer programs whose outputs are "*theirs*, not *ours*, with no human creator or director in sight."[275] That may well be the right way to understand generative AIs, either as they exist today or as they will exist soon. This Article's disagreement with Norton and Massaro is thus legal, not factual or philosophical. It is about whether being a truly independent speaker is, on its own, legally sufficient for an entity to claim First Amendment protections. And as shown above, under well-settled rules about the Constitution's limited scope, the answer is no.[276] Even assuming that AI outputs are speech, they are not *protected* speech.

Happily, the First Amendment's doctrinal toolkit is already equipped to handle facts exactly like these. As noted above, speech produced by unprotected speakers is extraordinarily common. Nearly all human speakers on planet Earth lack First Amendment rights.[277] For example, the Supreme Court has held that "[i]t is clear that" Belgian professors, as non-Americans outside the United States have "no constitutional right" to contest even the straightforward censorship of their speech.[278]

Nonetheless, Americans who *do* have First Amendment rights may sometimes desire to listen to the unprotected speech of Belgian professors.[279] In these circumstances, the First Amendment does in fact protect the Americans' interests.[280] But here, the interest is purely in *listening*, not *speaking*.[281]

Most of the scholars to have written about generative AI outputs and the First Amendment have suggested such "listeners' rights" as an important framework.[282] But none has thoroughly examined what *pure* listeners' rights consist of, shorn from the right to speak.

Often, listeners' rights are mentioned in the same breath as speakers' interests, suggesting parity between the two.[283] This is understandable. In cases involving *both* a protected speaker and a protected listener, the protections are often hard to distinguish. After all, if the government were free to plug the ears of every listener who wished to hear protected speech,

---

274. Or, possibly, the training data that produced them.
275. Massaro & Norton, *supra* note 4, at 1172.
276. *See supra* Section III.C.
277. Agency for Int'l Dev. v. All. for Open Soc'y Int'l, Inc., 591 U.S. 430, 433 (2020) (denying a First Amendment claim because "foreign citizens outside U.S. territory do not possess rights under the U.S. Constitution").
278. Kleindienst v. Mandel, 408 U.S. 753, 762 (1972).
279. *Id.*
280. *Id.* at 762–63.
281. *Id.*
282. *See supra* notes 1–7.
283. *Id.*

that would make a mockery of the protections for speech. The right to speak straightforwardly implies the right to be heard by willing listeners.[284] Thus, in cases where *both* speakers and listeners have First Amendment rights, constitutional protections for listening can be fairly strong.[285] This pattern characterizes most "listeners' rights" cases.[286]

But in those rarer cases where the speaker lacks the First Amendment right to communicate, all that remains are pure First Amendment protections for listening, qua listening. Those defenses are substantially weaker than the ones reserved for protected speaking, or even mixed acts of protected speaking and listening.

Commercial speech protections are sometimes said to be purely about listeners', rather than speakers', rights.[287] That is, advertisements are putatively protected not because they contain valuable expressive content, but for the sole sake of the consumer's interest in the "free flow of [commercial] information."[288] Commensurate with this listening-only theory of commercial speech, regulations of such speech are, at least in theory, subject to deferential First Amendment review.[289] However, as the Court has recognized, it is simply not true that advertisements cannot contain valuable expressive content—including political content.[290] Thus, in practice, commercial speech often implicates both protected speaking and protected listening, garnering more searching First Amendment protection.[291] Moreover, AI outputs are not, in general, advertisements. Thus, while theoretically informative, the commercial speech cases are not the best lens to analyze the listeners' rights implicated by regulations of AI outputs.

Kleindienst v. Mandel[292] presents a cleaner example of pure listening. There, the aforementioned Belgian professor—a Marxist—was invited to speak at Stanford.[293] The Supreme Court held that the professor, as a foreigner, was not a First Amendment protected speaker.[294] The Court refused to impute the professor's speech to the American academics who

284. Stanley v. Georgia, 394 U.S. 557, 564 (1969) ("[F]reedom of speech . . . necessarily protects the right to receive [speech]." (quoting Martin v. City of Struthers, 319 U.S. 141, 143 (1943)).

285. *See, e.g.*, Gregory v. City of Chicago, 394 U.S. 111 (1969).

286. *See, e.g.*, Martin v. City of Struthers, 319 U.S. 141 (1943); Thomas v. Collins, 323 U.S. 516 (1945).

287. *See* First Nat'l Bank v. Belotti, 435 U.S. 765, 784 (1978).

288. *Id.* at 784 & n.30.

289. Cent. Hudson Gas & Electric Corp. v. Pub. Serv. Comm'n, 447 U.S. 557 (1980).

290. Bigelow v. Virginia, 421 U.S. 809, 818 (1975) (discussing an advertisement for abortion services); *see also* N.Y. Times Co. v. Sullivan, 376 U. S. 254, 266 (1964).

291. *See, e.g.*, Sorrell v. IMS Health, Inc., 564 U.S. 552 (2011).

292. 408 U.S. 753 (1972).

293. *Id.* at 757.

294. *Id.* at 762.

invited him.[295] Thus, no protected speech, qua speech, was at issue. "The case, therefore, c[ame] down to the narrow issue" of what protections the "First Amendment confer[red] upon the appellee professors, because they wish[ed] to hear."[296]

What protections *does* the First Amendment confer for pure hearing? Specifically, what constitutional test applies to restrictions on listening to unprotected speech? *Kleindienst* does not offer a multipart test, just a few simple principles. The Belgian professor was denied entry to the United States under a statute forbidding visas to those who "advocate the . . . doctrines of world communism."[297] In a different constitutional context—one where the speaker had First Amendment rights—this law would, at a minimum, be a content-based restriction reviewed under strict scrutiny.[298] It would probably even be considered a viewpoint-based restriction and thus flatly unconstitutional.[299]

But the *Kleindienst* Court upheld the statute. And it did not apply strict scrutiny. Far from it. The Court simply asserted Congress's interest in, and control over, setting "[p]olicies pertaining to the entry of aliens and their right to remain here" and held that to be sufficient.[300] A good regulatory reason alone, *Kleindienst* suggests, is enough to overcome pure listeners' rights. How good, exactly? *Kleindienst* does not supply a single clear statement of the standard, but it does suggest descriptors ranging from "legitimate and bona fide" to "weight[y]."[301]

These relatively slim[302] protections for pure listening to unprotected speech make sense from the perspective of First Amendment theory. The reasons have already been discussed. For the same reason as the *production* of Belgian speech is of diminished First Amendment value, so is its *consumption*, even by Americans.[303]

Sunstein argues that *Kleindienst*'s immigration setting, not free speech principles, drove its outcome. In his view, the immigration context renders

---

295. *Id.* at 764.

296. *Id.* at 762.

297. *Id.* at 755 (citation omitted).

298. *See* Reed v. Town of Gilbert, 576 U.S. 155, 171 (2015).

299. *See* Matal v. Tam, 582 U.S. 218, 223 (2017).

300. *Kleindienst*, 408 U.S. at 766–67 (quoting Galvan v. Press, 347 U.S. 522 (1954)) (internal quotation marks omitted).

301. *Id.* at 770; *id.* at 764 (citation omitted).

302. Other First Amendment interests in consuming information, it is worth noting, are even weaker. The First Amendment does protect a pure right to receive information, even when no identifiable speaker, even an unprotected one, is attempting to supply it. But in such cases, even viewpoint-based restrictions are often explicitly allowed, as long as they are not "narrowly partisan or political." Bd. of Educ. v. Pico, 457 U.S. 853, 870 (1982). *But see* Zemel v. Rusk, 381 U.S. 1 (1965) (favoring a reasonableness standard).

303. *See supra* Section III.B. Indeed, the protections for American listeners are stronger than for Belgian speakers.

Kleindienst "exceedingly narrow," and "even unique," with little to say about listeners' rights in general.[304]

But that cannot be right. The nearly identical case of *Bridges v. Wixon* shows why.[305] *Bridges* was also an immigration case. Just like *Kleindienst*, it centered around a non-U.S.-citizen who published communist literature and whom the U.S. government therefore sought to exclude from the United States. But while *Kleindienst*'s Marxist was in Belgium, the defendant in *Bridges*, a union organizer, was already in the United States.[306] And, as the Court wrote, "Freedom of speech . . . is accorded aliens residing in this country."[307] As a result, unlike in *Kleindienst*, "the utterances made by [the organizer] were entitled" to treatment as First-Amendment-protected speech.[308] The result? The government could not deport the organizer unless it proved he advocated "overthrowing the government by force and violence."[309] This is a precursor to the *Brandenburg* test, which ranks among the most stringent First Amendment standards.[310]

Both *Kleindienst* and *Bridges*, then, were about "the entry of aliens and their right to remain" in the United States.[311] If Sunstein were right that *Kleindienst*'s lax First Amendment protections flowed from its immigration setting, then both cases would have been lax. But they were not. The difference between them—the one that explains their constitutional divergence—is that that *Bridges* involved a protected speaker, and thus protected speech.[312] *Kleindienst* involved only protected listening.

Pell v. Procunier supplies further evidence that the speaking–listening dichotomy explains *Kleindienst*.[313] It again deals with protected listening, isolated from protected speech. And it has nothing to do with immigration. In *Pell,* prisoners and journalists challenged a California law allowing prison officials to refuse face-to-face interviews with particular prisoners.[314] If such a ban on speaking to the press were imposed on members of the general public, it would likely have been treated as a presumptively unconstitutional prior restraint.[315] But here, the relevant speakers—

304. Sunstein, *supra* note 7, at 11.

305. 326 U.S. 135 (1945).

306. *Id.* at 137.

307. *Id.* at 148.

308. *Id.* at 148.

309. *Id.* The Court framed its holding as construing statutory language to avoid constitutional problems, rather than directly analyzing the statute's constitutionality. Such constitutional avoidance analyses, however, still reveal what the Court believed the Constitution required.

310. *Brandenburg* adds an imminence requirement. *See supra* Section II.A.

311. Kleindienst v. Mandel, 408 U.S. 763, 766–67 (1972).

312. Agency for Int'l Dev. v. All. for Open Soc'y Int'l, Inc., 591 U.S. 430, 434 (2020) (noting that "foreign citizens in the United States may enjoy certain constitutional rights" (emphasis removed)).

313. Pell v. Procunier, 417 U.S. 817 (1974).

314. *Id.* at 819.

315. *See* N.Y. Times Co. v. United States, 403 U.S. 713, 714 (1971).

imprisoned persons—lacked the First Amendment's full protections. As with other constitutional rights, a duly convicted incarcerated person loses those "First Amendment rights . . . inconsistent with his status as a prisoner."[316] This includes losing, the *Pell* Court determined, the right to speak to the media face-to-face.[317] *Pell* therefore involved, like *Kleindienst*, a ban affecting only unprotected speech.

But, as in *Kleindienst*, it also involved protected listening—the journalists' desire to hear what the prisoners had to say. Here, as in *Kleindienst,* the First Amendment applied, but review was quite deferential. Yet again, the Court upheld the law based just on the government's "substantial" justification for it.[318] Namely, avoiding turning specific prisoners into "virtual 'public figures'" whose "disproportionate degree of notoriety and influence [in the prison] . . . often became the source of severe disciplinary problems."[319] The Court raised no questions about narrow tailoring. Nor compelling interests. Nor imminence of harm. Nor intent. Here again, a strong interest, but not necessarily a compelling one, was sufficient to impinge on pure listeners' rights.[320]

If AI outputs are best understood as being like the speech of an unprotected speaker, this same deferential standard should apply to their regulation. That is, AI outputs should be regulable if the government supplies a bona fide, non-pretextual, and legitimate justification for the regulation.

This picture of AI outputs as unprotected speech is also factually attractive. To begin, it tracks a valid intuition underpinning much of the mistaken scholarly consensus view: namely, that AI outputs are extremely speechlike. They are full of ideas. They can be complex, funny, thoughtful, creative, and even profound. But unlike the constitutional models advocated in the emerging scholarly consensus, this model fits how AIs are actually made and used. It avoids the fallacy that AIs' outputs are, somehow, really

316. *Pell*, 417 U.S. at 822.

317. *Id.* at 822–28. *Pell* is not identical to *Kleindienst*, in that American prisoners, unlike foreigners, retain substantial First Amendment interests. As such, the Court spent more time in analysis before deciding that the prisoners lacked a right to speak to the press face-to-face. *Id.* at 822–29. Nonetheless, once the Court determined that the prisoners had no right to speak in that way, the case became one about pure listeners' rights.

318. *Id.* at 829.

319. *Id.*

320. Two cases are worth distinguishing here. First, in *Procunier v. Martinez*, 416 U.S. 396 (1974), the Court held that prisoners retain their First Amendment right to send uncensored letters. *Id.* at 415–16. Thus, while facially similar to *Pell*, *Martinez* involved both protected speech and protected listening. Second, *Lamont v. Postmaster General*, 381 U.S. 301, 302 (1965), involved a requirement that Americans register before receiving "communist political propaganda" by mail from abroad. Thus, while facially similar to *Kleindienst*, *Lamont* was primarily about the government's ability to force political dissidents to "out" themselves. *Id.* at 307. It is thus best understood as dealing with privacy and compelled speech, not pure listening. *Cf.* NAACP v. Alabama, 357 U.S. 449 (1958).

vessels for the ideas of human creators and users. If it is right to think of AI outputs as very much like speech, but wrong to think about them as human speech, only one option remains. AI outputs must then be, if anyone's expressions, the expressions of AIs themselves.

One recently posted whitepaper, by Karl M. Manheim and Jeffery Atik, strongly opposes the idea that AI outputs are expressions of any kind. "AI outputs are not speech in the first place," they say, so "[t]he First Amendment is beside the point."[321]

It is quite difficult to see how Manheim and Atik can be so sure of these claims, even by their own lights. Their core argument is that, because generative AIs "merely" learn to produce outputs that "statistically resemble[]" their training data, "[t]here are no ideas, thoughts, views or other creative elements involved."[322] Real speech in Manheim and Atik's view, requires at least some of these.[323] Surely, however, it can't require all of them, or else, for example, protesters chanting common political slogans would be classed as non-speakers, for lack of creativity.

Every part of Manheim and Atik's argument is, at a minimum, both contestable and hotly contested. First, the invocation of "mere" statistics recalls Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell's characterization of LLMs as "stochastic parrots" that lack "understanding."[324] The strongest version of that claim has now been proven false by empirical evaluations of modern LLMs. LLMs do not merely remix and regurgitate their training data. Instead, they abstract from training data, building various higher order generalizations—"understanding" is at least a good metaphor—and apply those generalizations to new examples.[325] As a result, LLMs can, for example, perform mathematical calculations, represent spaces, and deploy theory of mind to answer novel questions not included in their training data.[326] A pure "parrot" could not do any of that.

---

321. KARL MANHEIM & JEFFERY ATIK, WHITE PAPER: AI OUTPUTS AND THE FIRST AMENDMENT 4 (2023), https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID4527134_code332621.pdf?abstractid=4524263&mirid=1&type=2 [https://perma.cc/5ZR7-GCA7].

322. *Id.* at 3.

323. *Id.*

324. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, *in* FACCT '21 PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 610, 615, 617 (2021).

325. Yonathan Arbel & David A. Hoffman, *Generative Interpretation*, 99 N.Y.U. L. REV. 451, 476–77 (2024).

326. *See* OpenAI, *GPT-4 Technical Report* 5 (Mar. 27, 2023), https://cdn.openai.com/papers/gpt-4.pdf [https://perma.cc/Z8EM-PJWF] (table of GPT-4's performance on academic and professional exams); Sébastien Bubeck et al., Sparks of Artificial General Intelligence: Early Experiments with GPT-4 49–60 (Apr. 13, 2023) (unpublished manuscript), https://arxiv.org/pdf/2303.12712.pdf [https://perma.cc/25N3-RQPF].

Weaker versions of these claims are harder to evaluate. Perhaps Manheim and Atik do not mean that LLMs lack complex generalizations or the ability to solve puzzles they have not seen before. Perhaps they instead mean that LLMs lack *other* mental attributes—like consciousness—that generally accompany human "thoughts" and "ideas." Here, though, vast theoretical uncertainty precludes straightforward conclusions about either current AI systems or near-future ones. Neuroscience is nowhere close to fully characterizing the physical or biological processes necessary for consciousness.[327] Some prominent researchers doubt such characterization is possible, even in principle.[328] Others have argued that consciousness is an illusion—that, even for humans, it does not really exist.[329] Given all of this, it is doubtful that anyone, Manheim and Atik included, could say with confidence whether GPT-4 (or GPT-10, for that matter) had subjective mental experiences.

This is a law paper, though. And law deals better in pragmatics than metaphysics. As Wittes emphasizes, the outputs of generative AI have many of the practical features of speech: "They write text, they have dialogue with humans. They express opinions—however much they are incapable of believing anything."[330] Thus, people with First Amendment rights get the same benefits from consuming such expressions as they get from listening to the unprotected speech of genuine human speakers. Like a foreign academic, a generative AI can pose scientific hypotheses for confirmation or refutation. It can generate political claims, which might contribute to an electorate's informed decisionmaking. And it can assist protected speakers in developing views that they then express in acts of human autonomy. These pragmatic, rather than philosophical, factors supply good constitutional reason to treat AI outputs like speech—albeit the speech of a speaker lacking constitutional rights.

Thus, little depends on the question of whether AIs, existing or soon to exist, can "really" speak, whatever that would mean. The model treating AI outputs as a certain kind of speech works just as well analogically as literally. Even if AI outputs fail, in some sense, to be the literal equivalent of a Belgian Marxist's speech, they are a close match in all of the ways that the First Amendment says matters. Analogy being the lifeblood of legal reasoning, that is enough to warrant similar constitutional treatment.

---

327. *See generally* Patrick Butlin et al., Consciousness in Artificial Intelligence: Insights from the Science of Consciousness (2023) (unpublished manuscript), https://arxiv.org/pdf/2308.08708.pdf [https://perma.cc/Q2RR-CASG].

328. Thomas Nagel, *What Is It Like to Be a Bat?*, 83 PHIL. REV. 435, 435 (1974) ("[T]he mind-body problem [is] really intractable.").

329. Daniel C. Dennett, *Illusionism as the Obvious Default Theory of Consciousness*, 23 J. CONSCIOUSNESS STUD. 65 (2016).

330. Wittes, *supra* note 1.

2. *Tools for Speech*

Here is another legally workable and factually appealing way to think about the outputs of generative AIs. Not as AIs' own speech, nor as anyone else's, but as a tool or medium for producing or conveying speech.[331] This model captures especially nicely a set of AI use cases that one could think of as being at opposite ends of a spectrum—the most mundane and the most creative. In the middle of that spectrum are most of the generative AI uses discussed above. Things like a request for an LLM to write a poem or a diffusion model to generate an image. At the far mundane end of AI uses, however, are things like proofreading and editing. And on the creative end are, for example, various intersections of generative AI and performance art.

These AI use cases fit nicely into a set of existing First Amendment doctrines for dealing with other tools for speaking. As with listening, these tools are not themselves speech. But they can be highly useful to speakers, and they thus receive some First Amendment protection. Camping is not ordinarily communication. But camping en masse in a public park can be a powerful symbolic expression of support of the "plight of the homeless."[332] Likewise, a regulation against fires is not a direct regulation of speech. But burning a draft card can be an effective way to protest a war.[333] A loudspeaker is not speech. But it is a useful tool for, among other things, transmitting speech.[334] Likewise for the cables that carry television signals.[335] Likewise, too, for paper and ink. Indeed, these (and their digital analogues) are useful tools not just for transmitting ideas, but for composing them in the first place.

Generative AI systems, and their outputs, can be very useful to speakers in the same ways. They can perform some of the same functions that ink and paper, and digital word processors, can perform. Users can, for example, paste the entirety of an essay into ChatGPT and ask the model to reproduce the text verbatim, but with spelling, grammar, and punctuation errors corrected. AI systems can also be, like a free association exercise performed with pen and paper, wonderful tools for brainstorming ideas. The AI can

---

331. Volokh, Lemley, and Henderson briefly raise a similar view. *See supra* note 6 at 657–59. However, they contend that, as a tool for speaking, AI outputs are "fully protected by the First Amendment," like "the Internet and social media." *Id.* This Article explicitly rejects that model, advocating intermediate, rather than full, protection for AI outputs, and contrasting those protections with the stronger ones that social media posts receive. *See supra* notes 28–30 and accompanying text; *supra* notes 221–22 and accompanying text.

332. Clark v. Cmty. for Creative Non-Violence, 468 U.S. 288, 291–92 (1984).

333. United States v. O'Brien, 391 U.S. 367, 376 (1968).

334. *See* Ward v. Rock Against Racism, 491 U.S. 781, 784 (1989).

335. Turner Broad. Sys., Inc. v. FCC, 512 U.S. 622, 645 (1994) (applying intermediate scrutiny to a law imposed "based . . . upon the manner in which speakers transmit their messages").

suggest a range of possible words, phrases, or outlines, and the user can either reject these or adopt them as their own.

Such mundane uses may soon be the most common applications of generative AI. Both Microsoft's Office Suite and Google's Workspace applications now feature LLM integrations.[336] These integrations both augment existing outlining and editing functionality and greatly enhance suggested text and autocompletion. Once these integrations are widely adopted, the share of AI-generated words supplied for such mundane purposes may dwarf those supplied elsewhere. After all, the majority of words humans write—emails, memoranda, text messages—are similarly mundane.

More creatively, one can imagine uses of generative AI systems wherein the systems themselves, with their self-generated outputs—become a medium of expression. Suppose that, commenting on technology's increasing control over humanity, a performance artist wears a hidden earpiece via which an LLM can both hear and respond. For a month, the artist speaks only words supplied to her by the AI, not her own. Here, the AI outputs do not express the artist's thoughts—indeed, that is the point of the performance. But just as camping can, under certain conditions, become an expressive act, the AI system itself here becomes a kind of medium for symbolic expression.

When the government regulates a tool for producing or transmitting speech, or a medium for symbolically expressing it, these are called either "incidental" or "time, place, and manner" restrictions. The line between these two First Amendment categories is a blurry one. For example, the Court in *United States v. O'Brien* treated a prohibition on the burning of draft cards as "incidental" to the government's goal of preserving official documents.[337] But one could just as easily call it a prohibition on one "manner" of speaking—as the Court has said of prohibitions on sleeping in parks.[338]

Happily, however, the Supreme Court has written that the applicable tests are "little, if any, different."[339] Both amount, roughly, to intermediate scrutiny. The precise formulation of the test varies case to case. But passing it usually requires that the law serve a "substantial," "important," or "strong" government interest and be at least somewhat "tailored" to that

---

336. Jared Spataro, *Introducing Microsoft 365 Copilot – Your Copilot for Work*, OFF. MICROSOFT BLOG (Mar. 16, 2023), https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/ [https://perma.cc/2NEX-NHE3]; Johanna Voolich Wright, *A New Era for AI and Google Workspace* (Mar. 14, 2023), https://workspace.google.com/blog/product-announcements/generative-ai [https://perma.cc/X48K-P4JT].

337. 391 U.S. 367, 376 (1968).

338. Clark v. Cmty. for Creative Non-Violence, 468 U.S. 288, 297 (1984).

339. *Id.* at 298.

interest.[340] The asserted interest must be "unrelated to the suppression of free speech."[341] Tailoring is adequate if it avoids the situation where "a substantial portion of the burden on speech does not serve to advance [the law's] goals."[342] But the government's methods "need not be the least restrictive or least intrusive means of" achieving its goals.[343] And in any case, the regulatory regime the government chooses must "leave open ample alternative channels for communication."[344]

Insofar as generative AIs prove useful to protected speakers as tools or mediums for speech, regulations of their outputs are properly subjected to these moderate First Amendment standards.

IV. NON-SPEECH ANALYSES APPLIED

Now that we understand what generative AI outputs are, from the perspective of the First Amendment, we can return to the question of their regulation. Part II described the significant challenges AI safety regulations would face if subjected to the constitutional rules safeguarding protected speech. How, then, do safety laws' prospects change when the correct, less demanding, constitutional tests are applied? Nothing is certain at this point. The laws' yet-unwritten details might matter a great deal. Nonetheless, it is possible to identify, in broad strokes, the easy terrain and potential pitfalls for each type of safety law described in Part I. These observations will also provide valuable guidance to lawmakers in drafting regulations that will be upheld as constitutional—a necessary condition for their effectiveness.

A. Regulations of Dangerous Outputs

The lion's share of the necessary regulations described in Part I would be designed to safeguard against AI outputs that would directly threaten life and limb. Think here of rules designed to penalize outputs that would aid in the design of biological or chemical weapons, assist in the planning and execution of cyberattacks, or allow agentic autonomous systems to pursue objectives at odds with human safety. When protected speech poses dangers of this kind, regulations penalizing it are often subject to the *Brandenburg* rule.[345] That is, the government may punish the speech only if it can show

340. Geoffrey R. Stone, *Content-Neutral Restrictions*, 54 U. CHI. L. REV. 46, 48–49; *see also, e.g.*, City of Renton v. Playtime Theatres, 475 U.S. 41, 50 (1986) ("substantial"); *O'Brien*, 391 U.S. at 377 ("important or substantial"); Village of Schaumburg v. Citizens for a Better Env't, 444 U.S. 620, 636 (1980) ("strong").

341. *Clark*, 468 U.S. at 294.

342. Ward v. Rock Against Racism, 491 U.S. 781, 799 (1989).

343. *Id.* at 798.

344. *Clark*, 468 U.S. at 293.

345. *See supra* Section II.A.

that the speech was "directed to inciting or producing imminent lawless action" and "likely" to do so.[346] For AI outputs, this would be a very difficult showing to make.

But if AI outputs are better understood as unprotected speech, implicating only protected listening, the story changes dramatically. Then, *Kleindienst* suggests that Congress could forbid a wide range of outputs for the sake of preventing legitimate danger. Under *Kleindienst*, a legitimate regulatory interest supplies sufficient constitutional justification for interfering with protected listening to a wide range of unprotected speech. Indeed, in *Kleindienst*, the restriction was both content-based and directed at a particular political position. It was also extremely broad, burdening *all* speech "advocat[ing] the . . . doctrines of world communism," not just instances of speech advocating lawless action.[347]

Restrictions on dangerous AI outputs would be far less constitutionally fraught than those upheld in *Kleindienst*. The regulatory goal of preventing chemical and biological terrorism, cyberattacks, and rogue AI disasters would be similarly legitimate. But the target would be far less troublesome. The regulated outputs would be easily specified without any reference to politics at all. And they would be more closely tethered to their regulatory goal. Think, for example, of regulations penalizing potentially dangerous outputs discovered during pre-release testing—well before the danger becomes imminent. Thus, even if one views *Kleindienst* as being an especially regulation-friendly listeners' rights decision, AI safety laws should survive. At a minimum, they serve no less important a purpose, and are no more censorious, than the prison interview rules upheld in *Pell*.

The story is similar if AI outputs are treated as tools for speech, rather than speech—protected or otherwise. Here, *City of Renton v. Playtime Theatres, Inc.*[348] is instructive. That case involved a zoning law that restricted adult theaters to certain quarters of a city, for the sake of preventing crime.[349] Such a law is, in a literal sense, content based. It picks out the regulated theaters with reference to the kind of speech they exhibit. So too for AI safety regulations. The only way to know whether a given output contains an attempt to hack a power plant is to read it.

But *City of Renton* shows that such references to content do not necessarily trigger the most stringent constitutional tests. Specifically, they do not do so when the object of regulation is a tool for speaking rather than speech itself. Then, what matters is not whether the law *refers* to some speech's content, but rather whether it is "*justified* without reference to the

346. Brandenburg v. Ohio, 395 U.S. 444, 447 (1969).
347. Kleindienst v. Mandel, 408 U.S. 753, 755 (1972) (citations omitted).
348. 475 U.S. 41 (1986).
349. *Id.* at 44, 48.

content of the regulated speech."[350] In *City of Renton*, the Court held, the point of the law was not to regulate certain ideas. It was only to regulate the tangible harms—crimes—that exhibitions of certain kinds of content were, the Court thought, empirically likely to cause.[351]

So too for regulations of dangerous AI outputs. As with physical buildings, generative AI systems are useful tools for generating and refining human speech. Most of the speech users wish to generate using them is not likely to aid or cause tangible injuries like crimes. But some of it is. AI outputs that include instructions for chemical and biological weapons, contain plans for cyberattacks, or set artificial systems off on arbitrary and uncontrollable courses of action are particularly likely to cause such harms. Limits on such outputs would, like the zoning law in *City of Renton*, pick out what is forbidden, in part, by reference to content. But, as in *City of Renton*, the AI safety laws would not be "*justified*" on the basis of content. The idea is not, for example, that the ideological advocacy of bioterrorism is bad. It is that mass death from disease is bad, and certain outputs from highly capable AI systems are likely to help cause mass death.

Regulations of dangerous outputs would, of course, also have to conform with intermediate scrutiny's other requirements. They would have to serve a "substantial" government interest. Valid examples of "substantial" interests include preventing petty crime and reducing noise pollution.[352] Preventing mass homicide and the proliferation of uncontrolled, unaligned autonomous systems should count, too. The laws would also have to be at least moderately well-tailored to serving those goals.[353] To avoid overburdening speech, regulators should not, for example, forbid outputs that relate to chemistry in general. Instead, they should focus on limiting outputs going beyond education to constitute material assistance in the synthesis of especially dangerous compounds.[354]

Finally, to pass constitutional muster, regulations of dangerous outputs would have to "leave open ample alternative channels for communication."[355] This should pose little difficulty, since AI outputs rarely, if ever, instantiate human communications.[356] Insofar as they are

---

350. *Id.* at 48 (quoting Va. State Bd. Of Pharmacy v. Va. Citizens Consumer Council, Inc., 425 U.S. 748, 771 (1976)).

351. Note that it was far from empirically certain in *City of Renton* that adult theaters actually caused crime. This is a valid ground for attacking the Court's decision. But it goes to the specific facts of that case, rather than to the underlying constitutional principle.

352. *City of Renton*, 475 U.S. at 50; Ward v. Rock Against Racism, 491 U.S. 781, 796 (1989).

353. *Ward*, 491 U.S. at 798.

354. *See* Holder v. Humanitarian Law Project, 561 U.S. 1, 28 (2010) (favoring regulation aimed at prohibiting "material support" of terrorism, even when expressive, over regulation aimed at suppressing "pure political speech").

355. Clark v. Cmty. for Creative Non-Violence, 468 U.S. 288, 293 (1984).

356. *See supra* Section III.A.

useful tools for composing, refining, or transmitting human ideas, AI safety regulations would leave available all other such tools—word processors, social media, radio waves, and so on—that currently exist. Indeed, they would leave AI systems themselves available as tools for composing speech that did not directly threaten life and limb. That is to say, almost all speech.

B. Regulations of False and Deceptive Outputs

If safety regulations succeeded at preventing AI systems from producing outputs that directly threatened life and limb, that would be an immense success. But it would not complete the project of making powerful AI systems safe. Threats from false and deceptive outputs—socially engineered espionage, mass personalized propaganda—would still loom large. Regulations against such outputs will be needed, too. Much of the same First Amendment analysis applies here as applied to regulations of dangerous outputs. Rather than risk monotony, both this section and the next one focus on differences.

Fifteen years ago, one might have argued that regulating false AI outputs would be very easy, because the Supreme Court had often said that false speech was "valueless."[357] Thus, the Court sometimes said, such speech was "not protected by the First Amendment" at all.[358]

But as shown in Part II, that was never really true. Cases like *Gertz* and *Sullivan*, for example, require showings of a culpable state of mind to impose defamation liability on protected speakers, even for speech that is demonstrably false.[359] The reason is that speech containing falsities may also contain valuable commentary, advocacy, or expression. The substantial First Amendment protections afforded false statements uttered by protected speakers are thus designed to avoid "chilling" such intermingled high-value elements.[360]

But what if the speaker of the false speech lacks constitutional rights? Then the constitutional need for strong safeguards against chilling valuable elements intermingled with false ones falls away. The government is simply *allowed* to chill the speech of speakers without First Amendment rights. Not because their speech lacks value. But because that value, whatever it may be, falls outside the U.S. Constitution's scope of protection.

Thus, if AI outputs are understood as speech, but the speech of an unprotected speaker, safety regulations targeting falsity and deception

357. Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 52 (1988).
358. Brown v. Hartlage, 456 U.S. 45, 60–61 (1982).
359. N.Y. Times Co. v Sullivan, 376 U.S. 254, 279–80 (1964); Gertz v. Robert Welch, Inc., 418 U.S. 323, 347 (1974).
360. *Hustler*, 485 U.S. at 52; United States v. Alvarez, 567 U.S. 709, 723 (2012).

should, indeed, have an easy time. Here, the theoretical justifications for unconstrained regulations are, if anything, stronger than in the case of dangerous speech. Recall that, in this context, the First Amendment protects only an interest in listening. And the primary First Amendment value such listeners can receive from hearing protected speech is aid in the search for truth.[361] Dangerous speech may often contain true, and perhaps novel, ideas. Both a claim about a chemical formula for super-VX and a tenet of "world communism" may be true.[362] And even then, cases like *Kleindienst* and *Pell* allow substantial regulation based on a bona fide, legitimate regulatory interest. But false and deceptive speech, by definition, is *not* very useful in the search for truth. At best, it can serve as a foil for sincerely truth-seeking discourse. But it is not itself in the business of enlightenment. Usually, just the opposite.

Thus, under the standard of *Kleindienst* and *Pell*, AI safety regulations targeting false outputs should survive with relative ease. The regulatory goal of preventing deception is certainly "*bona fide*." Legal prohibitions on fraud and defamation stretch back centuries.[363] And the harms that could flow from widespread AI-assisted deception are quite important,[364] ranging from election disruption to espionage to the crippling of secure infrastructure. So long as lawmakers are evenhanded, regulating outputs that are legitimately false, as opposed to ideologically disfavored, the "listeners' rights" framework should raise few First Amendment difficulties for safety laws.

Likewise, if AI outputs are best understood as tools for speaking. As with dangerous outputs, the regulation of false outputs is, in one sense, literally content based. You need to know what something says to know whether it is true. And as with dangerousness, if protected speech were the object of regulation, this would raise a high constitutional hurdle—probably the *Sullivan* and *Gertz* rules. But, as above, when the government regulates only tools for or inputs to speech, being literally content based does not invoke the higher standard. Instead, what matters is the law's *justification*.[365] And as just discussed, the justifications for preventing falsity and deception are extremely well-established. Indeed, both legally and empirically, the harms from deception probably stand on firmer ground than the speculative public safety harms the Court endorsed in *City of Renton*.

---

361. As argued above, the speech of, say, a Belgian professor is neither an autonomous expression of an American rightsholder, nor part of the project of American democratic self-governance. It of course might be an *input* into both, but then it serves First Amendment values only as a second-order matter.

362. *See* Kleindienst v. Mandel, 408 U.S. 753, 755.

363. *See generally* Harry Cendrowski & Louis W. Petro, *History of Fraud Deterrence*, *in* THE HANDBOOK OF FRAUD DETERRENCE 15, 15–28 (2012); John C. Lassiter, *Defamation of Peers: The Rise and Decline of the Action for* Scandalum Magnatum, *1497–1773*, 22 AM. J. LEGAL HIST. 216 (1978).

364. *Alvarez*, 567 U.S. at 718, also rejects regulating lies to prevent trivial harms.

365. City of Renton v. Playtime Theatres, Inc., 475 U.S. 41, 48 (1986).

From there, the analysis is the same. To satisfy intermediate scrutiny's tailoring requirement, lawmakers should avoid overbreadth. They should stick to policing factual accuracy, not orthodoxy of opinion. And they should focus on the risks from AI specifically, leaving open those many alternative avenues and tools for human speech that presently exist.

C. Regulations of Racist and Bigoted Outputs

Anti-discrimination regulations for frontier AI will, like regulations of dangerous and deceptive outputs, be vital as AI systems become capable. If capable, biased AI systems are put in charge of complex institutions or organizations, they will further tilt those institutions against already disadvantaged groups. And even neutral, but highly compliant, systems could be deployed by bigoted humans to autonomously harass, intimidate, or even physically harm members of vulnerable populations.

Regulations designed to suppress racist and bigoted outputs raise more First Amendment complexities than those targeting dangerousness and deception. The complexities relate to the laws' differing aims. It is black-letter law that governments may legitimately regulate harms associated with threats, "fighting words" (extremely provocative or injurious insults), and physical violence.[366]

But as with regulating deception, when the object of regulation is the speech of a protected speaker, additional First Amendment safeguards fall into place. Then, the state may not *single out* racialized fighting words or threats for special prohibition because of their racialized character.[367] The idea is again that, while such speech contains some valueless elements, it implicitly contains others that the First Amendment treats as valuable. Here, the valuable elements are statements of (abhorrent) viewpoints on a (wrongfully) contested political question about the moral equality of racial groups.[368] Thus, a law singling out racialized threats because of their racialized character is "viewpoint discrimination" against a protected speaker entitled to express their repellent views.[369] Viewpoint discrimination is generally anathema to the First Amendment. It is almost always unconstitutional.[370]

---

366. R.A.V. v. City of St. Paul, 505 U.S. 377, 381 (1992); Beauharnais v. Illinois, 343 U.S. 250, 256 (1952); Wisconsin v. Mitchell, 508 U.S. 476, 484 (1993); Brandenburg v. Ohio, 395 U.S. 444, 447–48 (1969).

367. *R.A.V.*, 505 U.S. at 391–92; Virginia v. Black, 538 U.S. 343, 364–67 (2003) (rejecting a law for presuming that cross burning was, categorically, an unusually threatening act).

368. *R.A.V.*, 505 U.S. at 388–92.

369. *Id.* at 388.

370. Matal v. Tam, 582 U.S. 218, 234 (2017); Texas v. Johnson, 491 U.S. 397, 414 (1989).

The question is whether the problem of viewpoint discrimination persists when the object of regulation is something other than speech produced by a protected speaker. There are some reasons to think it does not. Recall that the statute upheld in *Kleindienst* singled out a specific political ideology for special prohibition. That is, for the sake of national security, it burdened communist, but not, for example, anarchist speech. This supports the idea that, if AI outputs are best understood as being like foreign speech, harmful and racialized outputs could likewise be specially prohibited.

To be clear, even here, there is little reason to think that a viewpoint restriction could be justified purely by the odiousness of the views expressed. Even in *Kleindienst*, the asserted goal was national security. And the viewpoint singled out was the one most associated with the era's most imminent national security threat—the Soviet Union. Discriminatory speech, and discriminatory AI outputs, can likewise cause concrete harms. Consider, for example, the straightforward economic costs of workplace sexual harassment.[371] The important question is whether, for the sake of preventing such concrete, non-ideological harms, the government may single out expressions reflecting a particular contested viewpoint. *Kleindienst* suggests that the answer might be yes.

However, this may be where Sunstein's insistence on *Kleindienst*'s exceptionality is most plausible. Cases like *Bridges* and *Pell* convincingly show that regulations affecting only protected listening to completely unprotected speech are reviewed quite deferentially. A legitimate and bona fide goal is usually sufficient.[372] But only *Kleindienst* itself went further, upholding a viewpoint-based restriction. Maybe that final leap requires something special. Perhaps here is where the confluence of immigration, national security, and unprotected speakers matters. After all, even when pure protected listening is the only First Amendment activity at stake, viewpoint-based restrictions seem especially pernicious. It is one thing to restrict listening to unprotected speech relating to some general topic. It is another to systematically warp the unprotected speech on a topic to which protected listeners have access, in service of political orthodoxy.

Thus, even if AI outputs are rightly understood as the speech of an unprotected speaker, the constitutionality of laws forbidding bigoted outputs is uncertain. If *Kleindienst* is read for all it is worth, such viewpoint-based rules might survive. But if *Kleindienst* involved unusual government discretion, and *Pell* presents the ordinary framework for regulating pure listening, they may not.

---

371. *See* Hishon v. King & Spalding, 467 U.S. 69, 78 (1984); Meritor Savings Bank, FSB v. Vinson, 477 U.S. 57, 65 (1986).

372. *See supra* note 300 and accompanying text.

The model of AI outputs as a mere tool or medium for speech is more promising. Here the case law is clearer. In *Wisconsin v. Mitchell* the Supreme Court upheld a law singling out hate crimes for special punishment.[373] It distinguished that statute from laws prohibiting racial threats or fighting words on the ground that crimes—there, a beating—are not themselves speech.[374] They are instead conduct.[375]

Mitchell echoes *Renton* and the other tools-for-speaking cases throughout. As in those cases, the fact that a crime is not usually speech changed the First Amendment analysis.[376] But it did not end the inquiry. Instead, the Court credited the free speech dimension of a law punishing only crimes motivated by, and potentially communicating, certain ideas.[377] As in *Renton*, the law itself referred to the content of those ideas in defining the proscribed conduct. But as in *Renton*, that was not sufficient to trigger heightened scrutiny, much less the specter of viewpoint discrimination. What mattered, again, was not whether the law *referred* to content, but whether it was *justified* by it. In *Mitchell*, the law's singling out of "bias-motivated crimes" was justified by the claim that such crimes are more "destructive of the public safety and happiness."[378]

Much the same could be said of bigoted AI outputs. Especially if those outputs lead to concrete injuries—like exclusion from economic and social life. Victims of such discrimination have more cause for emotional outrage than victims of non-racialized erroneous exclusion.[379] Moreover, a discriminatory AI system can do much more societal harm than a single biased human. A human's malign influence is naturally circumscribed by her decisional purview and limited capacity for work. A discriminatory AI can produce injustice at scale. This, combined with an ample literature showing that, absent intervention, AIs are likely to reproduce bias in their training data, is worrying. It supplies significant justification for treating biased AI outputs as a zone of special regulatory concern.

Nonetheless, cautious AI safety regulators may wish to avoid writing special rules for racist and bigoted outputs. This is in part because they could prevent a large share of racial injustice without such rules. Recall that the major constitutional concern with singling out racism is one of viewpoint discrimination. This means that, paradoxically, *broader* prohibitions are less constitutionally fraught. For example, even in the fighting words and threats

373. 508 U.S. 476 (1993).
374. *Id.* at 487.
375. *Id.*
376. *Id.* at 484–85.
377. *Id.*
378. *Id.* at 488.
379. *Id.* (citing increased "emotional harms" as a legitimate justification for punishing hate crimes).

contexts—where protected speech is clearly at issue—statutes punishing *all* threats and fighting words are allowed.[380] Such prohibitions would of course also cover racialized threats and fighting words.

Similarly, AI safety regulators might require that powerful AI systems refrain from threats or fighting words of all kinds. They might require that AI-generated decisions are fair and accurate along a wide array of important dimensions, not just race and other protected statuses. If such regulations are well-enforced, compliance will have to include racial, along with other kinds of, fairness.

There may of course be general reasons to prefer targeted remedies for racial injustice over generalized ones. Among them, such laws' symbolic value and their ability to constrain the use of limited enforcement resources. But there is a trade-off at stake. A targeted law that is struck down as unconstitutional accomplishes much less than a generalized law that is upheld.

CONCLUSION

Generative AI is a transformative technology. It holds great promise across the range of human endeavors. It could help us make new discoveries,[381] cure old diseases,[382] accelerate economic growth, and even lift billions out of poverty. But like all new technology, increasing power increases both reward and risk. AI catastrophes of many kinds therefore loom. They can be avoided. But only if governments succeed in implementing effective safety requirements—and technologists then succeed in making the breakthroughs needed to implement them.

According to emerging scholarly theories, the First Amendment will pose a serious threat to such innovations. Those theories advocate treating AI outputs on the same footing as the protected speech of humans who bear constitutional rights. But the theories are mistaken. Both legally and factually, AI outputs are not best understood as being protected speech. Other models, already well-established in First Amendment theory and doctrine, are a better fit. Under these models, AI outputs would receive non-trivial First Amendment protections. They could not be forbidden by wildly overbroad laws imposed to serve unimportant regulatory goals. But under these constitutional standards—the ones guarding protected listening and tools for speaking—most well-crafted AI safety regulations should easily

380. R.A.V. v. City of St. Paul, 505 U.S. 377, 393–94 (1992); Counterman v. Colorado, 600 U.S. 66, 69 (2023).

381. *See, e.g.*, Bernardino Romera-Paredes et al., *Mathematical Discoveries from Program Search with Large Language Models*, 625 NATURE 468 (2024).

382. *See, e.g.*, Liu et al., *supra* note 15.

survive. These nuanced understandings, both of First Amendment doctrine and of generative AI, will be vital to the legal project of making AI safe.