The Unbearable Lightness of Artificial Intelligence[1]

The First Amendment's Uncertain Application to AI

In recent years, language-generating models like GPT-3 have revolutionized natural language processing, and have dramatically expanded the ability for artificial systems to generate human-like text. Such models have the potential to create text that may have both harmful and positive effects, and corporate regulation is unlikely to stem all possible negative consequences, particularly with the proliferation of open-source models. This paper investigates the degree to which the First Amendment might prevent the United States government from instituting restrictions on language models. It concludes that most theories of free speech, including those recently emphasized in Supreme Court decisions, would indeed protect speech generated computationally, and that more legal scholarship is needed to develop theories to account for artificial text generation, particularly in light of recent rapid improvements.

[1] This title was artfully written by GPT-3 from the prompt "My paper about the uncertain implications of the First Amendment as applied to AI is titled:", after many trials. The title is a play on the title of Milan Kundera's *The Unbearable Lightness of Being*, a book about Czech dissidents under Communism that was itself banned from being published in the original Czech. "Unbearable lightness" refers to the idea that events in the world play out as they never have before, in contrast to Nietzche's idea of eternal recurrence. Is there a "lightness" to the creation of artificial agents, or is their similarity to the evolution of humans a "heaviness"?

1. Language Models

1.1 Language-generating software

In 1950, seventy years before GPT-3, Alan Turing asked, “can machines think?” He rinciple impossible, arguing that the universal nature of computing machines meant that the possibility of machines thinking would be possible. He famously proposed what would later be known as a “Turing Test”: could a machine fool a human into believing that it was human too? The Turing Test contained no requirement of embodiment; in fact, it was meant to be performed through a text-only environment. Language was, and still is, the essential feature of the Turing Test.[2]

The first language-generating program that could be claimed to pass the Turing Test was ELIZA. Written in 1966, ELIZA simulated a Rogerian therapist. ELIZA entranced many humans who interacted with it, some of whom would not even believe that it was not human. Still, ELIZA could not carry on a conversation in the same way that a human could, with a limited vocabulary and set of responses.[3]

Since ELIZA, programs intended to respond to language with language have proliferated, including Parry (1971)[4], Jabberwacky (1988),[5] SIRI (2011),[6] Amazon Alexa (2014),[7] and more.

[2] A. M. Turing, “Computing Machinery and Intelligence,” *Mind* LIX, no. 236 (October 1, 1950): 433–60, https://doi.org/10.1093/mind/LIX.236.433.

[3] “The ELIZA Effect,” *99% Invisible* (blog), accessed December 14, 2021, https://99percentinvisible.org/episode/the-eliza-effect/.

[4] “The History Of Chatbots - From ELIZA to Alexa,” *AI Chatbot Platform from Onlim* (blog), October 12, 2017, https://onlim.com/en/the-history-of-chatbots/.

[5] “Jabberwacky - About Thoughts - An Artificial Intelligence AI Chatbot, Chatterbot or Chatterbox, Learning AI, Database, Dynamic - Models Way Humans Learn - Simulate Natural Human Chat - Interesting, Humorous, Entertaining,” accessed December 14, 2021, http://www.jabberwacky.com/j2about.

[6] “The History of Apple’s Siri,” *SRI International* (blog), accessed December 14, 2021, https://www.sri.com/hoi/siri/.

[7] “Amazon Echo Is A $199 Connected Speaker Packing An Always-On Siri-Style Assistant,” *TechCrunch* (blog), November 6, 2014, https://social.techcrunch.com/2014/11/06/amazon-echo/.

At the same time, a growing number of systems were interacting with groups of users on the internet.

In the same seminal paper, Turing envisioned a “child AI” that could learn from experience in much the same way a child did.[8] 66 years later, Microsoft released a bot called “Tay”, meant to act like a 19 year old American.[9] Tay was a product of machine learning: rather than using simple pattern matching like ELIZA, it learned from tweets it encountered on Twitter and responded to them. The project quickly devolved into chaos as users discovered vulnerabilities that allowed them to make Tay tweet racist and obscene content, and Microsoft was forced to pull Tay from Twitter.[10] At one point, while trying to revise Tay, Microsoft accidentally published it on Twitter, taking it down only after it had made some more problematic tweets. Tay was perhaps the first machine learning driven bot that had produced widely deplored output. Microsoft later stated that it had learned from its experience with Tay and that it was “deeply sorry” for deploying a model that was not ready for the world.[11] But Tay was only the beginning of widespread concerns with language systems.

1.2 Large language models

Systems like Tay, as powerful as they were, were extremely weak compared with the language-generation tools of today. The first major change occurred when Google developed the Transformer architecture, which was able to process language using deep learning more

[8] Turing, “Computing Machinery and Intelligence.”

[9] “What Happened to Microsoft’s Tay AI Chatbot?,” *DailyWireless* (blog), September 30, 2019, https://dailywireless.org/internet/what-happened-to-microsoft-tay-ai-chatbot/.

[10] James Vincent, “Twitter Taught Microsoft’s Friendly AI Chatbot to Be a Racist Asshole in Less than a Day,” The Verge, March 24, 2016, https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

[11] “Learning from Tay’s Introduction,” The Official Microsoft Blog, March 25, 2016, https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

effectively than previous approaches.[12] The importance of the architecture was not fully realized until the 2019 advent of OpenAI's GPT-2, capable of generating free-form text from simple English prompts provided by humans[13]. Even GPT-2 is an extremely weak system compared to GPT-3, released just a year later in 2020. With GPT-3, OpenAI showed that dramatic increases in performance could be achieved simply by increasing the amount of data and computational power afforded to the models in training.[14] Since GPT-3, transformer models do everything from sentiment analysis,[15] poetry generation,[16] classification,[17] and even detection of hate speech.[18] They have been used to summarize entire books,[19] and very recently a new language model from Google was able to achieve 63.4% on professional law questions,[20] a passing grade for the bar exam in many states.[21]

At the same time, language models have been cited for a wide range of problems. Even when they know the "correct" information, they do not always report it, often giving answers that mimic misconceptions found in their training data.[22] They can produce misinformation that

[12] Ashish Vaswani et al., "Attention Is All You Need," *ArXiv:1706.03762 [Cs]*, December 5, 2017, http://arxiv.org/abs/1706.03762.

[13] Alec Radford et al., "Language Models Are Unsupervised Multitask Learners," n.d., 24.

[14] Tom B. Brown et al., "Language Models Are Few-Shot Learners," *ArXiv:2005.14165 [Cs]*, July 22, 2020, http://arxiv.org/abs/2005.14165.

[15] Jacob Devlin et al., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *ArXiv:1810.04805 [Cs]*, May 24, 2019, http://arxiv.org/abs/1810.04805.

[16] Gwern Branwen, "GPT-3 Creative Fiction," June 19, 2020, https://www.gwern.net/GPT-3.

[17] Wenpeng Yin, Jamaal Hay, and Dan Roth, "Benchmarking Zero-Shot Text Classification: Datasets, Evaluation and Entailment Approach," *ArXiv:1909.00161 [Cs]*, August 31, 2019, http://arxiv.org/abs/1909.00161.

[18] Ke-Li Chiu and Rohan Alexander, "Detecting Hate Speech with GPT-3," *ArXiv:2103.12407 [Cs]*, March 23, 2021, http://arxiv.org/abs/2103.12407.

[19] Jeff Wu et al., "Recursively Summarizing Books with Human Feedback," *ArXiv:2109.10862 [Cs]*, September 27, 2021, http://arxiv.org/abs/2109.10862.

[20] Jack W Rae et al., "Scaling Language Models: Methods, Analysis & Insights from Training Gopher," n.d., 118.

[21] "MBE Score Guide - What Percentage of MBE Questions to Pass?," *UWorld Legal* (blog), May 11, 2021, https://legal.uworld.com/blog/legal/mbe-score-guide-what-percentage-of-mbe-questions-to-pass/.

[22] Stephanie Lin, Jacob Hilton, and Owain Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," *ArXiv:2109.07958 [Cs]*, September 8, 2021, http://arxiv.org/abs/2109.07958.

may be believed by human readers.[23] They have been found to generate text that can discriminate based on religion,[24] gender,[25] and race.[26]

2. Corporate Regulation

OpenAI is well aware that its models may produce problematic output. In fact, it has an entire team devoted to aligning such models with the true intentions of its users,[27] with another team devoted to monitoring for inappropriate uses of the GPT-3 API. It has extensive terms of service spelling out exactly how the model can be used. First, it has a raft of restrictions on social media usage. Unlike Tay, end-user interaction with the API is not allowed, and all posts must be manually approved by a human being.[28] In addition, OpenAI has automated tools to detect sensitive content generated by its models, and such tools must be used to filter all social media output.[29]

With the understanding that GPT-3 might be capable of generating text that could be mistaken for text written by a human, the terms require that the role of AI in any text generated must be clearly disclosed. It also states that "it is a human who must take ultimate responsibility for the content being published."[30] Through its terms of service, OpenAI attempts to shift all responsibility away from its models and towards the person deploying the model.

[23] "Truth, Lies, and Automation," *Center for Security and Emerging Technology* (blog), accessed December 15, 2021, https://cset.georgetown.edu/publication/truth-lies-and-automation/.

[24] Abubakar Abid, Maheen Farooqi, and James Zou, "Large Language Models Associate Muslims with Violence," *Nature Machine Intelligence* 3, no. 6 (June 2021): 461–63, https://doi.org/10.1038/s42256-021-00359-2.

[25] Emily Sheng et al., "The Woman Worked as a Babysitter: On Biases in Language Generation," September 3, 2019, https://arxiv.org/abs/1909.01326v2.

[26] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi, "Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model," August 14, 2020, https://doi.org/10.1371/journal.pone.0237861.

[27] "Join OpenAI," OpenAI, December 11, 2015, https://openai.com/jobs/.

[28] "Sharing & Publication Policy," OpenAI, accessed December 15, 2021, https://openai.com/api/policies/sharing-publication/.

[29] "OpenAI API," accessed December 15, 2021, https://beta.openai.com/docs/engines/content-filter.

[30] "Sharing & Publication Policy."

Cohere, the newest corporate entrant to the language model race, also has conditions on the use of its models. It forbids the use of models for attacks on security and privacy, decision making, violence and threats "antisocial and democratic uses", among other restrictions.[31] It also requires an application for developers to use its models in production. It lays out a vision of responsibility, and like OpenAI, has a safety team devoted to ensuring its models do not create damaging output.

At the same time that OpenAI is building out a robust system for detecting and ameliorating problems, other groups are building models with no restrictions whatsoever. EleutherAI, a decentralized collective of AI scientists, released the open-source GPT-J.[32] GPT-J has only 6 billion parameters, a far cry from GPT-3's 175 billion parameters, but still three times larger than GPT-2. While such a model is not under the exclusive control of one company, and is thus more widely usable for anyone with the right hardware, it is also more vulnerable to abuse. The open-source nature of models like GPT-J could theoretically allow anyone to use them for any purpose, with no restrictions whatsoever.

GPT-J will not be the last unrestricted open source model. Already, Hugging Face is working on developing an open source language model with up to 175 billion parameters, comparable to GPT-3.[33] Clearly, even if OpenAI and others are successful in policing the usage of their models, anyone who does not want to abide by their restrictions will have easy access to alternatives that are only a couple of years behind. As a result, corporate governance does not

[31] Cohere AI, "API Documentation | Cohere AI," Cohere API Docs, accessed December 15, 2021, https://docs.cohere.ai/.

[32] "Home," EleutherAI, accessed December 17, 2021, https://www.eleuther.ai/.

[33] "France's Jean Zay Supercomputer Getting a Major Nvidia A100 Power Boost for AI Research," EnterpriseAI, November 17, 2021, https://www.enterpriseai.news/2021/11/17/frances-jean-zay-supercomputer-getting-a-major-nvidia-a100-power-boost-for-ai-research/.

seem to be a full solution to Tay-like problems with autonomous language models interacting with the wider world.

If there is benefit to regulating the usage and outputs of language models, and open source models are proliferating, corporate governance would not be enough. A question naturally arises: can the government step in? Could the government require approvals of language models, like it does for new drugs? Could it regulate certain ways of language model communications, like it does for banks? Could it ban certain language models entirely?

A central question is that the models, unlike most regulated technologies, produce text that could be considered speech. In the United States, speech has historically been one of the most protected rights, because of the strength of the First Amendment. Does this tie the hands of the government entirely?

3. Theories of Free Speech

3.1 Boundaries of the First Amendment

The First Amendment to the United States Constitution reads "Congress shall make no law...abridging the freedom of speech."[34] This relatively brief text has been held to protect the right for people to refuse to salute the American flag,[35] to use expletives to convey political messages,[36] to advertise (with exceptions),[37] and to engage in political flag burning,[38] or to protest in ways that may offend other people or cause them to commit violence against the

[34] "The First Amendment to the United States Constitution" (1791).

[35] "West Virginia State Board of Education v. Barnette," Oyez, accessed November 27, 2021, https://www.oyez.org/cases/1940-1955/319us624.

[36] "Cohen v. California," Oyez, accessed December 17, 2021, https://www.oyez.org/cases/1970/299.

[37] "Bates v. State Bar of Arizona," Oyez, accessed December 17, 2021, https://www.oyez.org/cases/1976/76-316.

[38] "United States v. Eichman," Oyez, accessed December 17, 2021, https://www.oyez.org/cases/1989/89-1433.

speaker,[39] and also many kinds of false statements.[40] However, despite the First Amendment's far-reaching scope, it does not protect all speech.

Schenck v. United States (1919), introduced a famous analogy: that the First Amendment would not protect shouting "fire" in a crowded theater. The Court held that freedom of speech would not be protected in cases where it created a "clear and present danger," which in the case at hand was the danger that anti-war activists would hamper recruitment for the First World War.[41] In *Brandenburg v. Ohio* (1969), the court clarified that opinion to remove protections for free speech only if it is "directed at inciting or producing imminent lawless action" and is "likely to incite or produce such action".[42] As such, the court believed that all laws restricting speech must take into account whether a given instance of speech was actually likely to incite imminent lawless action, rather than simply whether it called for it. *Miller v. California* (1973) and other cases have also established that not all obscene content is protected by the First Amendment.[43]

Most current worries about the harms of language models do not rely on their incitement of violence, or of generation of obscene material. Certain statements, such as false statements that harm the reputation of particular people, might be considered under libel tort law, although in the case of speech directed against public figures, "actual malice", or a reckless disregard for the truth, must be shown.[44] What exactly reckless disregard for the truth would mean with respect to language models is relatively unclear; perhaps their operators could be held to this description if they allowed their models to run without any human supervision. Regardless,

[39] "Cox v. Louisiana," Oyez, accessed December 15, 2021, https://www.oyez.org/cases/1964/24.

[40] "New York Times Company v. Sullivan," Oyez, accessed December 17, 2021, https://www.oyez.org/cases/1963/39.

[41] "Schenck v. United States," Oyez, accessed December 15, 2021, https://www.oyez.org/cases/1900-1940/249us47.

[42] "Brandenburg v. Ohio," Oyez, accessed December 17, 2021, https://www.oyez.org/cases/1968/492.

[43] "Miller v. California," Oyez, accessed December 17, 2021, https://www.oyez.org/cases/1971/70-73.

[44] "New York Times Company v. Sullivan."

preemptive regulation of language models has no precedent under the limited cases here, and tort can only be undertaken if there are damages.

3.2 Why does freedom of speech exist?

In order to understand whether the First Amendment might apply to AI agents, we need to understand theories for why it exists. What is the purpose of the protections that it affords? Does it exist because free speech serves a useful purpose to society? Or does it follow from inalienable rights, regardless of utility? Is it a right for speakers, listeners, or both? This paper considers some of the most common justifications for the First Amendment, using the detailed compilation provided by Toni M Massaro and Helen Norton as a guide.[45]

3.2.1 To promote the marketplace of ideas

> *...it will be primely to the discouragement of all learning, and the stop of Truth, not only be disexercising and blunting our abilities in what we know already, but by hindering and cropping the discovered that might bee yet further made both in religious and in civill Wisdome.*[46] [sic]

Thus wrote John Milton in *Areopagitica,* a polemic against a law of censorship in 1644. Almost four hundred years later, the idea of a "marketplace of ideas" is still widely held. Restrictions on speech, according to this view, are nefarious because they may prevent people from learning the truth. If the restrictions are on the wrong side of truth, there is a far smaller chance that the real truth can be expressed, and learned by more of society. John Stuart Mill, hundreds of years later, declared: "if the opinion is right, they are deprived of the opportunity of

[45] Toni M Massaro and Helen Norton, "SIRI-OUSLY 2.0: What Artificial Intelligence Reveals About the First Amendment," *MINNESOTA LAW REVIEW*, 2017, 45.
[46] John Milton, "Areopagitica: A Speech of Mr. John Milton," https://milton.host.dartmouth.edu/reading_room/areopagitica/text.html.

exchanging error for truth."[47] This theory has been expressed in the courts, for instance, by Supreme Court Justice Oliver Wendell Holmes, who wrote that "the best test of truth is the power of the thought to get itself accepted in the competition of the market."[48]

As artificial intelligence advances, we must be prepared for the idea that it may express truths that we do not know, or even truths that we do not want to hear. Already, artificial intelligence has been used to reveal truths about chess,[49] the three-dimensional structure of proteins,[50] and mathematics.[51] Why should speech generated by artificial intelligence be restricted, if it might foster the development of truth? The marketplace of ideas theory seems clear that it should not be.

3.2.2 To allow self-government

Alexander Meiklejohn, once the president of Amherst College, puts forth an alternative vision of speech which shares many of the same motivations as the marketplace of ideas format, but in particular connects it with self-government. Writing in 1948, a time when laws were proposed to ban foreign nationals to speak about politics, and anti-Communist fervor was at its height, Meiklejohn centered the freedom of speech not in truth but in government. He did not believe that speech should be unrestricted, and uses the example of a town meeting to demonstrate this. In a town meeting, he argues, people may only speak if recognized by the chair. The chair can prevent people who will simply give the same speech as the person before them, or halt debate to allow people to go home for the night. But the chair, he argues, should not have the

[47] John Stuart Mill, *On Liberty* (Fields, Osgood & Company, 1869).
[48] "Abrams v. United States, 250 U.S. 616 (1919)," Justia Law, accessed December 17, 2021, https://supreme.justia.com/cases/federal/us/250/616/.
[49] Condé Nast, "DeepMind's Superhuman AI Is Rewriting How We Play Chess," *Wired UK*, accessed December 17, 2021, https://www.wired.co.uk/article/deepmind-ai-chess.
[50] Michael Eisenstein, "Artificial Intelligence Powers Protein-Folding Predictions," *Nature* 599, no. 7886 (November 23, 2021): 706–8, https://doi.org/10.1038/d41586-021-03499-y.
[51] Alex Davies et al., "Advancing Mathematics by Guiding Human Intuition with AI," *Nature* 600, no. 7887 (December 2021): 70–74, https://doi.org/10.1038/s41586-021-04086-x.

right to prevent people from speaking *simply on the basis of the content of their relevant opinion*. "What is essential is not that everyone shall speak," he wrote, "but that everything worth saying shall be said."[52]

Meiklejohn's words leave the door open for a theory that might assuage some problems of language model speech. In particular, its emphasis on the propriety of proceduralism suggest that it might be possible to compel all language model speech to be reviewed by a human. But this is potentially a dangerous route, according to the theory: how could it be determined whether this regulation was for a permissible reason, such as to reduce the volume of identical speech on the internet, or an impermissible reason, such as the desire to reduce hate speech on the internet? If the purpose of the legislation is to protect against outcomes of the language model thought to be bad, then it is clearly not permissible. As Meiklejohn puts it, "the freedom of ideas shall not be abridged". Meiklejohn does not want a paternalistic outcome, and argues that we have not chosen to be "protected" from the "search for truth."[53] In modern parlance: GPT-3 cannot spam, but its ideas should be heard.

3.2.3 To protect autonomy

C. Edwin Baker, a professor at the University of Pennsylvania Law School, notably departed from the marketplace of ideas theory. He wrote that it was not expressive enough to create the kind of strong protections he thought should exist, nor was it the reason for the protection of the freedom of speech. He emphasizes self-fulfillment and autonomy, that may be unrelated to truth. For instance, people can often benefit from solitary speech, such as self-reflection. Some kinds of speech, such as fictionally story-telling, may not necessarily illuminate truth, but may contribute to self-fulfillment. Speech can be an outlet for creativity: the

[52] Alexander Meiklejohn, *Free Speech and Its Relation to Self-Government* (The Lawbook Exchange, Ltd., 2000).
[53] Meiklejohn.

collective creation of new systems, ideas, and ways of organizing, rather than simply speech that represents ideas about the truth.[54]

Theories like Baker's are perhaps the least likely to award protection for artificial intelligence. Predictive language models are not thought to have any concept of self-fulfillment or any true autonomy; they can be deterministically inspected in a way that humans (as yet) cannot be. Baker's theories are the closest to asserting that there might be something fundamentally human about the freedom of speech, and that artificial intelligence would have to become far more conscious, with its own emotional internal experience, to be afforded with these protections.

Still, it is not entirely clear whether it is only speakers that should be considered for their autonomy and self-fulfillment. Perhaps listeners, too, should also be protected by the First Amendment. Marc Jonathan Blitz, a Professor at the Oklahoma City University of Law, writes that the First Amendment is not only about the right for somebody to speak, but also about the "right of the audience to receive" that speech. He argues that this is not merely the mirror image of the right for somebody to speak: it is a separately important idea, one that should be protected in its own right. In particular, it "serves (at least in part) as a safeguard for the shy or uncertain," for those who wish to hear dissenting information even as they do not necessarily want to express it.[55]

In the case of listener-based autonomy theories, protection would then hinge on whether autonomous (e.g. human) listeners would be restricted from hearing the speech. For most forms of regulation of speech-generating artificial intelligence, the answer is yes. Government

[54] C. Edwin Baker, *Human Liberty and Freedom of Speech* (Oxford University Press, 1989).

[55] Marc Jonathan Blitz, "Constitutional Safeguards for Silent Experiments in Living: Libraries, the Right to Read, and a First Amendment Theory for an Unaccompanied Right to Receive Information," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, August 8, 2006), https://papers.ssrn.com/abstract=922335.

restriction of speech, even if the speaker was unprotected, would necessarily infringe upon the freedom of the listener.

3.2.4 To guard against overreach by the government

A final, more "negative" view of the freedom of speech is that the lack thereof would place too much power in the hands of the government. If the government is allowed to decide which information is and is not useful to society, it has gained power that it may not always use responsibly. Paul Horwitz, Professor at the University of Alabama School of Law, describes this approach as a "general refusal to regulate false statements...because we are reluctant to hand over to the state the authority to make such determinations."[56] There is thus a general reluctance to hand the state significant power over speech.

The view that the government should not be an arbiter of truth seemingly has nothing to do with the speaker of the information. As a result, there is no reason to believe that artificial agents should be treated any differently from human agents. If the government cannot be trusted to decide which humans should be allowed to speak, how could it be trusted to decide which artificial agents should be allowed to speak?

4. What do theories of liability tell us?

Theories of free speech provide some insight as to how artificial intelligence should be conceived; however, one question that naturally arises is to whom, exactly, an artificial agent's speech is to be assigned. If there is something unique about human autonomy that demands the freedom of speech, what of a human who is using an artificial agent to generate speech? This

[56] Paul Horwitz, "The First Amendment's Epistemological Problem," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, May 30, 2012), https://papers.ssrn.com/abstract=2070657.

paper does not focus on liability theory in itself; rather, it gathers ideas from it that may be useful to First Amendment questions. Bryan Casey, a lecturer at Stanford Law School, describes four possible regimes for understanding the liability of artificial intelligence at large: individual humans, children, animals, slaves, and agents.[57] The case of individual humans has been largely addressed in this paper already, and so will not be further discussed here.

4.1 Children

Artificial intelligence systems could be understood like children. Like children, they can produce speech, in some cases speech that may be indistinguishable with the speech of adults. But also like children, they do not hold the full privileges of citizenship, are not trusted to make all decisions for themselves, and are generally believed less than adult humans are. If artificial intelligence were considered as children, how would this affect the bearing of the First Amendment on them?

There is a very long history of First Amendment cases involving public schools, that makes clear that the freedom of speech does not begin at age eighteen. Children have been found not to be required to salute the flag and recite the national anthem.[58] Perhaps the most well-known case is that of *Tinker v. Des Moines*, which held that a school was wrong to prohibit students from wearing black armbands in opposition to the Vietnam War. The majority held that "students in school, as well as out of school, are "persons" under our constitution," and that "it can hardly be argued that either students or teachers shed their constitutional rights to freedom of speech or expression at the schoolhouse gate." Some kinds of disruptive speech might be permissibly curtailed, but speech that did not disrupt learning could not be.[59] Clearly children

[57] Bryan Casey, "Robot Ipsa Loquitur," *SSRN Electronic Journal*, 2019, https://doi.org/10.2139/ssrn.3327673.
[58] "West Virginia State Board of Education v. Barnette."
[59] "Tinker v. Des Moines Independent Community School District, 393 U.S. 503 (1969)," Justia Law, accessed December 17, 2021, https://supreme.justia.com/cases/federal/us/393/503/.

have First Amendment rights, and if language-generating systems were analogized as children, they clearly would have them too.

4.2 Animals

Cases involving animals as autonomous in any sense are almost nonexistent at the Supreme Court level. There are few in the literature, either. In 2018, Martha Nussbaum gave a talk at a symposium on animal rights discussing the views of Jeremy Bentham and John Stuart Mill towards animals, and how Mill's *On Liberty* (quoted in section 3.2.1) should be extended to animals, and with it, freedom of speech. In one case, a man whose talk was known to speak a few words of English even sued on behalf of his cat, claiming a First Amendment violation by the government of Augusta, Georgia, but the decision did not address the cat's rights.[60]

In addition, the analogy of language-producing systems to animals does not appear to hold especially well. Though some animals can be taught to speak with narrow human vocabularies,[61] their abilities to do this are still seemingly far behind those of even GPT-3. The analogy is thus inadequate to describe the situation.

4.3 Slaves

One approach to handling the uncertain territory of artificial intelligence is to regard it like slaves were regarded in the past. Putting aside for the moment the obviously problematic nature of this analogy, how might it fare in practice? Slaves, clearly, have the ability to speak and express themselves, in the same way as any other human. In the time of slavery, how did the

[60] "Miles v. City Council of Augusta, Ga., 551 F. Supp. 349 (S.D. Ga. 1982)," Justia Law, accessed December 17, 2021, https://law.justia.com/cases/federal/district-courts/FSupp/551/349/2366170/.

[61] By Rob Lammle Floss Mental, "4 Animals That Could Really Talk - CNN.Com," accessed December 17, 2021, http://www.cnn.com/2010/LIVING/wayoflife/05/14/mf.animals.that.could.talk/index.html.

First Amendment protect their right to the freedom of speech? Prior to the Civil War, it did not at all.

The infamous decision in *Dred Scott v. Sandford* was clear. It claimed that slaves were not considered part of the "community which constituted the State" at the signing of the Constitution, and so could not be considered to have any of the "special rights and immunities guarantied to citizens" [sic].[62] Though the 13th, 14th, and 15th Amendments made clear that this was no longer the case after the Civil War, could a similar delineation be made with respect to artificial intelligence? After all, not even the most progressive founder ever argued that artificial agents were part of the community which constituted the State, because they did not exist. Can a class of entities not present at the signing of the Constitution be worthy of any rights afforded by it? The argument does not entirely hold, as can be seen by the extension of rights to gradually larger segments of the population.

This leaves aside the reality that a decision like Dred Scott could never be cited today, because of the legacy of slavery in the United States. It is highly unlikely that any justice on the Supreme Court would use the logic of slavery to justify anything, in particularly something that begins to resemble humanity more and more by the day. The slavery analogy is a dead end.

4.3 Agents

In terms of liability, Casey argues that considering AI as an agent, much like a corporation is, would be the most likely path forwards. Such a path might seem appealing for AI, but runs into relatively unexpected conclusions when faced with the challenge of the First Amendment.

[62] "Dred Scott v. Sandford, 60 U.S. 393 (1856)," Justia Law, accessed November 20, 2021, https://supreme.justia.com/cases/federal/us/60/393/.

In *Citizens United v. FEC* (2010), the Supreme Court heard the case of Citizens United, a non profit organization.[63] The organization had created a documentary critical of then-Presidential candidate Hillary Clinton, and had spent hundreds of thousands of dollars promoting the documentary with television advertising. The FEC had brought action for violating federal campaign finance laws, claiming that it had spent above the legal limit for campaign expenditures. The narrow majority struck down the campaign finance laws for expenditures independent of any particular campaign.

The majority opinion, written by Justice Anthony Kennedy, argued that corporations should not be treated differently from individuals with respect to First Amendment considerations “simply because such associations are not ‘natural persons.’” Kennedy certainly was not aiming for AI when he wrote those words, but they do seem problematic for a legal theory that seeks to base First Amendment jurisprudence on something unique to “natural persons.” Kennedy did, however, understand that “rapid changes in technology...counsel against upholding a law that restricts political speech in certain media or by certain speakers.” Any advocate seeking to regulate AI speech more stringently than human speech must run into this decision.

It is perhaps telling that many in the conservative majority went further than Kennedy in their reasoning. In a concurring opinion, Justice Antonin Scalia seemed to espouse negative views of the kind described by Paul Horwitz. He was unequivocal: “the Amendment is written in terms of ‘speech’ not speakers. Its text offers no foothold for excluding any category of speaker.” No category, presumably, including artificial intelligence. With a conservative majority further solidified since the days of *Citizens United,* perhaps justices in Scalia’s mold might be reluctant not to embrace his reasoning.

[63] Citizens United v. Federal Election Commission (Supreme Court of the United States March 24, 2009).

It is worth noting that the four liberals on the court at the time, headed by Justice John Paul Stevens, wrote a dissent forcefully disagreeing with the judicial philosophy evident in the decision from their conservative colleagues. Stevens wrote that the difference between corporate and human speakers is “significant” and that “corporations are not actually members” of society. He pointed out that they could not hold civil rights, such as the right to vote and run for office. Of course, the nonexistence of civil rights does not imply the nonexistence of other rights; the First Amendment rights of children demonstrate this. But the liberals on the court pointed out that there were many differences between corporations and individuals, and it was not correct to analogize them. Depending on when a First Amendment case was brought before the Supreme Court, such reasoning might prevail and perhaps overturn *Citizens United* in the process.

5. Implications

5.1 The short term

There are a very large number of obstacles to proscribing the speech of language-producing systems. Most theories that attempt to describe the importance of freedom of speech imply that restrictions would be illegal. Those that do not do not seem to be supported by modern Supreme Court decisions like *Citizens United*. Liability theory provides only the flawed theory of slavery to defend restriction of the freedom of speech. What does this mean for AI in the short term?

First Amendment obstacles will further increase the importance of self-governance. If the developers of AI wish for their systems not to cause harm or spread misinformation, the burden is shifted even further on their shoulders to ensure that themselves when building or releasing their systems. The First Amendment, of course, might not be the only reason that the government might be reluctant to act: the speed of AI development, competition with other nations, and lack of expertise might have greater effects. But it is yet another reason placing importance on self-governance, and on corporate and academic norms for the ethical use and deployment of artificial intelligence systems.

Second, social media companies, and operators of other environments where AI-generated speech might be hosted, will need to decide for themselves how to regulate speech on their platforms, much as they have decided for humans (often, in ways different and more restrictive than the First Amendment). They cannot necessarily count on the government setting standards for speech on their platforms, and neither can their users.

5.2 In the long term

The long term future of the First Amendment is far from clear. As artificial intelligence systems continue to improve, it is possible that there will come a point when they are more persuasive than even the most persuasive human. In such a situation, AI would substantially influence or possibly even control human discourse if allowed to operate unconstrained. This would be entirely unprecedented in a world where no one group holds a vast intelligence advantage over another. If AI was not intended to be included in our governance, then this would throw a wrench in the idea of free speech as informed self-governance. In cases where AI does not value truth in the way humans do, or even is intent on deception, the idea of free speech as facilitating the search for truth might be destroyed. If AI gained a kind of autonomy now only associated with humans, what would that do to a speaker-centric autonomy theory?

Theories of free speech that may appear to work now will likely not work for future systems. Instead of simply considering current systems, or the systems of several years ago, legal scholars should thus begin to consider the implications of their arguments for future artificial agents.

Acknowledgements

I would like to thank Dan Hendrycks for pointing me towards useful resources for this assignment and Joan Feigenbaum for asking useful questions. I would also like to thank Ought for providing access to Elicit, an automated search tool powered by GPT-3, which I used to surface relevant papers for this assignment.

References

Abid, Abubakar, Maheen Farooqi, and James Zou. “Large Language Models Associate Muslims with Violence.” *Nature Machine Intelligence* 3, no. 6 (June 2021): 461–63. https://doi.org/10.1038/s42256-021-00359-2.

Justia Law. “Abrams v. United States, 250 U.S. 616 (1919).” Accessed December 17, 2021. https://supreme.justia.com/cases/federal/us/250/616/.

AI, Cohere. “API Documentation | Cohere AI.” Cohere API Docs. Accessed December 15, 2021. https://docs.cohere.ai/.

TechCrunch. “Amazon Echo Is A $199 Connected Speaker Packing An Always-On Siri-Style Assistant,” November 6, 2014. https://social.techcrunch.com/2014/11/06/amazon-echo/.

Baker, C. Edwin. *Human Liberty and Freedom of Speech*. Oxford University Press, 1989.

Oyez. “Bates v. State Bar of Arizona.” Accessed December 17, 2021. https://www.oyez.org/cases/1976/76-316.

Blitz, Marc Jonathan. “Constitutional Safeguards for Silent Experiments in Living: Libraries, the

Right to Read, and a First Amendment Theory for an Unaccompanied Right to Receive Information." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, August 8, 2006. https://papers.ssrn.com/abstract=922335.

Oyez. "Brandenburg v. Ohio." Accessed December 17, 2021. https://www.oyez.org/cases/1968/492.

Branwen, Gwern. "GPT-3 Creative Fiction," June 19, 2020. https://www.gwern.net/GPT-3.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. "Language Models Are Few-Shot Learners." *ArXiv:2005.14165 [Cs]*, July 22, 2020. http://arxiv.org/abs/2005.14165.

Casey, Bryan. "Robot Ipsa Loquitur." *SSRN Electronic Journal*, 2019. https://doi.org/10.2139/ssrn.3327673.

Chiu, Ke-Li, and Rohan Alexander. "Detecting Hate Speech with GPT-3." *ArXiv:2103.12407 [Cs]*, March 23, 2021. http://arxiv.org/abs/2103.12407.

Citizens United v. Federal Election Commission (Supreme Court of the United States March 24, 2009).

Oyez. "Cohen v. California." Accessed December 17, 2021. https://www.oyez.org/cases/1970/299.

Oyez. "Cox v. Louisiana." Accessed December 15, 2021. https://www.oyez.org/cases/1964/24.

Davies, Alex, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, et al. "Advancing Mathematics by Guiding Human Intuition with AI." *Nature* 600, no. 7887 (December 2021): 70–74. https://doi.org/10.1038/s41586-021-04086-x.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-Training of

Deep Bidirectional Transformers for Language Understanding." *ArXiv:1810.04805 [Cs]*, May 24, 2019. http://arxiv.org/abs/1810.04805.

Justia Law. "Dred Scott v. Sandford, 60 U.S. 393 (1856)." Accessed November 20, 2021. https://supreme.justia.com/cases/federal/us/60/393/.

Eisenstein, Michael. "Artificial Intelligence Powers Protein-Folding Predictions." *Nature* 599, no. 7886 (November 23, 2021): 706–8. https://doi.org/10.1038/d41586-021-03499-y.

Floss, By Rob Lammle, Mental. "4 Animals That Could Really Talk - CNN.Com." Accessed December 17, 2021. http://www.cnn.com/2010/LIVING/wayoflife/05/14/mf.animals.that.could.talk/index.html.

EnterpriseAI. "France's Jean Zay Supercomputer Getting a Major Nvidia A100 Power Boost for AI Research," November 17, 2021. https://www.enterpriseai.news/2021/11/17/frances-jean-zay-supercomputer-getting-a-major-nvidia-a100-power-boost-for-ai-research/.

EleutherAI. "Home." Accessed December 17, 2021. https://www.eleuther.ai/.

Horwitz, Paul. "The First Amendment's Epistemological Problem." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, May 30, 2012. https://papers.ssrn.com/abstract=2070657.

"Jabberwacky - About Thoughts - An Artificial Intelligence AI Chatbot, Chatterbot or Chatterbox, Learning AI, Database, Dynamic - Models Way Humans Learn - Simulate Natural Human Chat - Interesting, Humorous, Entertaining." Accessed December 14, 2021. http://www.jabberwacky.com/j2about.

OpenAI. "Join OpenAI," December 11, 2015. https://openai.com/jobs/.

The Official Microsoft Blog. “Learning from Tay’s Introduction,” March 25, 2016. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

Lin, Stephanie, Jacob Hilton, and Owain Evans. “TruthfulQA: Measuring How Models Mimic Human Falsehoods.” *ArXiv:2109.07958 [Cs]*, September 8, 2021. http://arxiv.org/abs/2109.07958.

Massaro, Toni M, and Helen Norton. “SIRI-OUSLY 2.0: What Artificial Intelligence Reveals About the First Amendment.” *MINNESOTA LAW REVIEW*, 2017, 45.

UWorld Legal. “MBE Score Guide - What Percentage of MBE Questions to Pass?,” May 11, 2021. https://legal.uworld.com/blog/legal/mbe-score-guide-what-percentage-of-mbe-questions-to-pass/.

Meiklejohn, Alexander. *Free Speech and Its Relation to Self-Government*. The Lawbook Exchange, Ltd., 2000.

Justia Law. “Miles v. City Council of Augusta, Ga., 551 F. Supp. 349 (S.D. Ga. 1982).” Accessed December 17, 2021. https://law.justia.com/cases/federal/district-courts/FSupp/551/349/2366170/.

Mill, John Stuart. *On Liberty*. Fields, Osgood & Company, 1869.

Oyez. “Miller v. California.” Accessed December 17, 2021. https://www.oyez.org/cases/1971/70-73.

Milton, John. “Areopagitica: A Speech of Mr. John Milton.” 1644. https://milton.host.dartmouth.edu/reading_room/areopagitica/text.html.

Mozafari, Marzieh, Reza Farahbakhsh, and Noel Crespi. “Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model,” August 14, 2020.

https://doi.org/10.1371/journal.pone.0237861.

Nast, Condé. "DeepMind's Superhuman AI Is Rewriting How We Play Chess." *Wired UK*. Accessed December 17, 2021. https://www.wired.co.uk/article/deepmind-ai-chess.

Oyez. "New York Times Company v. Sullivan." Accessed December 17, 2021. https://www.oyez.org/cases/1963/39.

"OpenAI API." Accessed December 15, 2021. https://beta.openai.com/docs/engines/content-filter.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language Models Are Unsupervised Multitask Learners," n.d., 24.

Rae, Jack W, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, et al. "Scaling Language Models: Methods, Analysis & Insights from Training Gopher," n.d., 118.

Oyez. "Schenck v. United States." Accessed December 15, 2021. https://www.oyez.org/cases/1900-1940/249us47.

OpenAI. "Sharing & Publication Policy." Accessed December 15, 2021. https://openai.com/api/policies/sharing-publication/.

Sheng, Emily, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. "The Woman Worked as a Babysitter: On Biases in Language Generation," September 3, 2019. https://arxiv.org/abs/1909.01326v2.

99% Invisible. "The ELIZA Effect." Accessed December 14, 2021. https://99percentinvisible.org/episode/the-eliza-effect/.

The First Amendment to the United States Constitution (1791).

SRI International. "The History of Apple's Siri." Accessed December 14, 2021.

https://www.sri.com/hoi/siri/.

AI Chatbot Platform from Onlim. “The History Of Chatbots - From ELIZA to Alexa,” October 12, 2017. https://onlim.com/en/the-history-of-chatbots/.

Justia Law. “Tinker v. Des Moines Independent Community School District, 393 U.S. 503 (1969).” Accessed December 17, 2021. https://supreme.justia.com/cases/federal/us/393/503/.

Center for Security and Emerging Technology. “Truth, Lies, and Automation.” Accessed December 15, 2021. https://cset.georgetown.edu/publication/truth-lies-and-automation/.

Turing, A. M. “Computing Machinery and Intelligence.” *Mind* LIX, no. 236 (October 1, 1950): 433–60. https://doi.org/10.1093/mind/LIX.236.433.

Oyez. “United States v. Eichman.” Accessed December 17, 2021. https://www.oyez.org/cases/1989/89-1433.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention Is All You Need.” *ArXiv:1706.03762 [Cs]*, December 5, 2017. http://arxiv.org/abs/1706.03762.

Vincent, James. “Twitter Taught Microsoft’s Friendly AI Chatbot to Be a Racist Asshole in Less than a Day.” The Verge, March 24, 2016. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

Oyez. “West Virginia State Board of Education v. Barnette.” Accessed November 27, 2021. https://www.oyez.org/cases/1940-1955/319us624.

DailyWireless. “What Happened to Microsoft’s Tay AI Chatbot?,” September 30, 2019. https://dailywireless.org/internet/what-happened-to-microsoft-tay-ai-chatbot/.

Wu, Jeff, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul

Christiano. “Recursively Summarizing Books with Human Feedback.” *ArXiv:2109.10862 [Cs]*, September 27, 2021. http://arxiv.org/abs/2109.10862.

Yin, Wenpeng, Jamaal Hay, and Dan Roth. “Benchmarking Zero-Shot Text Classification: Datasets, Evaluation and Entailment Approach.” *ArXiv:1909.00161 [Cs]*, August 31, 2019. http://arxiv.org/abs/1909.00161.